

Research assessment's ups and downs

The funding councils of the United Kingdom have achieved high standards in assessing, and thereby raising, the research performance of British universities. But dangers in the process merit further analysis.

SPARE a thought for the panellists involved in the third research assessment exercise (RAE) just carried out by British university funding councils (see *Nature* 385, 2; 1996). Carrying out the assessment of all university departments in their respective disciplines was burdensome enough. Now, returning to their desks after a seasonal break, they will have been assailed by telephone calls from friends and perhaps some freshly minted enemies seeking feedback on or justification for their decisions.

But in general the signs are that the panellists can reflect on a job well done. Cries of outrage have been understandably vigorous from those who feel their departments have been underrated — but those appear to be few in number. The process whereby up to four publications per staff member over a period of four years were judged alongside editorships, plenary talks and the like appears to have evolved to a point where the community perceives a fair trade-off between adequate refinement of the assessments and workloads required to achieve it. The scale of assessment has seven bands, from poor quality to international quality virtually throughout a department. Panellists acknowledge uncertainties at the margins but some, at least, assert that departments have been rated with a strong-confidence interval smaller than a band.

But there remain broader questions about the impact of the process. On the positive side, standards as a whole are improving. Over the four years since the previous exercise, universities have decided to invest in some departments to boost their ratings; staff have been pressurized to apply themselves more vigorously to achieving grant income and productivity. The RAE has also led to other tactics new to British academic life. Some high-fliers been bought in at high prices to boost some department performances while, more questionably, other stars have delayed departures for the same reason. No doubt some staff relations have deteriorated and there are allegations that teaching has suffered (although the evidence of the latter is slender, and it is worth remembering that universities are separately rated and rewarded for teaching too).

The pressures in universities are also strong in another sense. Higher education budgets have declined and the United Kingdom's new universities are putting additional pressure on the system. For those reasons, a university that has sustained its departmental rankings will nevertheless receive less money per staff member during the next funding round later this year.

Such stresses may be the price of progress in a time of declining budgets. But other financial impacts also need to be considered. The ratings will have a direct impact on money received for staff and infrastructure from the funding councils, although the formulae to be applied are not yet clear. The university receives the money as a block, so distribution to departments remains at the discretion of the institution (whereby hang some colourful tales of internecine strife). But the RAE has implications for other revenues: studentships from research councils have, on at least one occasion, been deliberately biased to highly rated departments; awards of research grants can also be influenced as, undoubtedly, will be a department's ability to attract industrial and other external support.

The RAE's strong leverage on ability to attract funds should therefore cause concern. It makes all the more questionable the policy of funding councils not to consider appeals against ratings.

But that leverage may also be introducing distortions, favouring those who dream of a superleague of research universities, and undermining a cause that supporters of the RAE tend to celebrate: the ability of newer or poorly endowed departments to improve themselves and thus breed healthy diversity. □

Rally behind the NSF

Scientists should support a proposal to expand the budget of the National Science Foundation by more than 7 per cent.

THE new year opens with some encouraging signs that the scientific community in the United States is waking up to its responsibility to provide more vocal and public support for the National Science Foundation (NSF), the main funding agency for non-bio-medical university research in the United States.

The American Chemical Society (ACS), the Federation of American Societies for Experimental Biology (FASEB), the American Physical Society and other groups will work together this spring to persuade the Congress of the importance of the agency (see *Nature* 384, 393; 1996).

FASEB and the ACS say NSF funding should grow by 7.1 per cent this year, restoring the 6.1 per cent in spending power which FASEB estimates the agency has lost in the past two years, and add a gentle expansion of another one per cent (see page 103). The physicists — pedantic to the last — point out that the second significant figure lacks an empirical basis.

Two years ago, FASEB was mocked for proposing a 10 per cent increase in the budget of the National Institutes of Health (NIH) at a time of national austerity. "Earth to FASEB..." ran the headline in *Science* (266, 1935; 1994). It turned out that FASEB knew something *Science* did not: in 1995 and again last year, the NIH attracted increases at and close to the FASEB target.

The NSF is not the NIH, and it will never be treated quite so fondly by the Congress. But the increment that the societies seek on its behalf is much smaller (\$232 million, against almost \$1 billion won last year for NIH), and there is a strong case to be put for it.

Both the Clinton administration and the Congress have spent the past two years preaching the special value of peer-reviewed, university science. They frequently cite the NSF as the model sponsor of such science, and invite other agencies to learn from its example. The administration, meanwhile, produces countless 'reviews' of its \$20 billion-a-year network of government laboratories, criticizing their wastefulness and excessive overhead costs. These criticisms are then redoubled at congressional hearings.

But neither branch of the government has mustered the courage to rationalize the laboratories and risk derailing the federal gravy train in even one congressional district. What they should do is to subject the laboratory network to real reform, and so release some resources to support the best available university research through the NSF. Clinton is not expected to do this in his budget proposal next month. The community must, therefore, persuade the Congress to choose peer-reviewed science over pork-barrel spending, and restore the NSF's spending power. □



Political will and cash 'needed to speed up removal of landmines'

Boston. New technologies and strategies could speed up removal of the more than 100 million landmines that remain buried around the world, according to two reports issued last month. But the initiatives need a greater injection of both political will and finance to succeed.

Landmines kill or maim 25,000 to 30,000 people a year, 80 per cent of them civilians, say the reports. There is a pressing need for a single high-level intergovernmental agency to focus global efforts on implementing new technologies for demining, says one report written by Kosta Tsipis. The author is a physicist at Massachusetts Institute of Technology (MIT), and the report is based on a 'brainstorming workshop' held last August in Cambridge, Massachusetts.

Paul Horowitz, a physicist at Harvard University, is lead author of the second report, *New Technological Approaches to Humanitarian Demining*. This is based on a study conducted by the JASON defence advisory group at the request of the Defense Advanced Research Projects Agency.

"Demining removes about 100,000 mines annually, but for every mine removed about 20 new mines are emplaced," says the JASON report. "The problem is only getting worse."

Hand-held metal detectors used for demining cannot distinguish between mines and other metal objects buried in the soil. So the task is perilous and time-consuming with false alarms outnumbering genuine detections, sometimes by 1,000 to one.

"People have become discouraged by the slow and dangerous nature of demining," says Tsipis. He organized the MIT work-

shop, and invited a group of scientists known for technical innovation, in the hopes of making the process safer, more efficient and less costly. "We came out of the workshop encouraged by the promising technologies that can be brought to bear on the problem."

The first step, he says, is to determine the precise characteristics of landmines to make detection easier. "We still do not know the physical and chemical properties of landmines." Measurements are needed to establish their acoustic, microwave and infrared 'signatures'.

A combination of new technologies could speed up demining by a factor of five, according to the MIT report. These are a metal detector called the "interdigitated winding magnetometer"; an "air knife" that can dig up mines in seconds, and the explosive Lexfoam which can blow up mines in situ. Tsipis is thinking of organizing a workshop this year to see how these devices might be used in Laos to clear the millions of unexploded 'bomblets' dropped during the Vietnam War.

Both reports recommend various methods to help to identify mines, including nuclear quadrupole resonance, X-ray backscatter techniques and electronic vapour detectors. But such technologies need further development.

Within five years, detectors mounted on robotic vehicles could enhance demining capabilities 100- to 1,000-fold, the MIT report predicts. But it adds that no single approach will solve all demining problems.

Although the JASON study was confined to technology assessment, Tsipis and his colleagues also addressed political concerns.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Unwelcome legacy: better demining methods could reduce the number of landmine victims.

Tsipis argues that a central agency — whether the US government, United Nations or European Union — is required to coordinate efforts and raise and allocate funds for new technologies. Such an agency would rely on a science advisory board to provide new information and to serve as a link between academics and industry.

"Today, there are dozens of nongovernmental organizations (NGOs) like Red Cross and Care working separately, which doesn't get you anywhere," Tsipis says.

The call for a central agency is not likely to receive a universal welcome. One US defence official says: "There already is a central agency — the Humanitarian Demining Technologies program run by the Department of Defense (DoD), which has tested more than 30 pieces of demining equipment since it began in 1995." He also disputes the claim that money is the chief stumbling-block: "Technology is the problem, not money."

Tsipis praises the DoD programme for its contributions to technology development and testing. But he adds that a higher-level agency is needed "to make decisions about what technologies will be used and where the money will come from". The United States has appropriated \$20 million to the task for 1997, and this may be enough for research and development, he says. "But the crunch will come when you need hundreds of millions of dollars for demining on a large scale."

Steve Nadis

BSE inquiry could prompt censure vote

Paris. The European Parliament may be asked to bring a motion of censure against the European Commission over its handling of the bovine spongiform encephalopathy (BSE) crisis. The use of this power would be unprecedented in the history of the European Union, and could result in the dismissal of individual commissioners.

The censure motion is proposed in a draft report of the conclusions of the parliament's inquiry into the crisis by Manuel Medina Ortega, a Spanish socialist member of the parliament, and rapporteur of the inquiry. The report alleges that the commission had a policy of "minimizing the problem [of BSE] that

could have given rise to a policy of disinfection, always with the aim of preventing perturbation of the beef market".

This is one of a long list of allegations in the 26-page report of "negligence" against the commission, the Council of Ministers and the UK government — many of which had already been made during the course of the inquiry (see *Nature* **384**, 8; 1996 & **385**, 6; 1997). In particular, the report strongly criticizes the failure of the UK government to ban exports of potentially contaminated meat and bone meal after it had decided to ban its use in ruminant feed at home.

The parliament's inquiry is due to end later this month.

Declan Butler

UK studies grapple with school science ...

Birmingham, England. Teachers and education researchers in Britain are preparing to carry out a top-to-bottom review of the school science curriculum in an attempt to address the question: what is the purpose of teaching science?

Two parallel consultation exercises are already under way to canvass ideas about a future science curriculum. In one, the Association for Science Education (ASE) is analysing the responses of many of its 20,000 members under an initiative called 'Science Education 2000+'.

A second project, 'Science Education for the Future', organized by the School of Education at King's College, London, and the Nuffield Foundation, was announced at the ASE's three-day annual meeting in Birmingham last week.

Science education in Britain has been reviewed several times over the past decade. It has been affected by a number of factors, including the revamping of the school-leaving examination taken at 16, compulsory tests during earlier years, a new national curriculum with specific learning targets and the introduction of compulsory science for primary schoolchildren.

But, despite the changes, "there has never been any real questioning of who science education is meant to serve", says Rosalind Driver, professor of education at King's College. There is a perception that present science curricula focus on acquiring facts — considered essential for a career in research — but not enough on the skills needed in non-research careers: analysis, investigation and a conscious awareness of the process of science.

The ASE consultation, the results of

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Recipe for science? Two projects ask what science should be taught, and to whom.

which are due to be published in the summer of 1998, asks fundamental questions, such as "what is science?", "why do science?" and "what science should be taught?". The King's College study, which is due to be completed at the end of 1998, will focus on the outcome of a number of detailed seminars and open meetings.

"The world of science is changing," says Mary Ratcliffe, chair of the ASE and lecturer in science education at the University of Southampton. "There is a lot of controversy — look at xenotransplants, for example — and more exposure to science in the making, such as life on Mars.

"We should be trying to prepare people to deal with these issues. Everyone does the national curriculum. But we now need to ask if everyone needs to do the same type of physics, chemistry and biology."

Examination statistics show a decline in the proportion of pupils specializing in the sciences after the age of 16. Yet most of those who reply to pupil surveys consider

the impact of science on their lives to be "very important", says Jonathan Osborne, a lecturer in science education at King's College. "This enigma needs to be resolved. "We need to ask what kind of science education will you design for the majority. Will there be one type, or more? And can we afford not to educate pupils to the highest levels of science and technology?"

The present curriculum, Osborne says, emphasizes 'classical' science. "But science increasingly is about uncertainty and risk assessment", and involves politics and economics. "Just look at issues such as BSE." Osborne adds that people need to understand the machinery of science, and how decisions are made that involve the interaction of science and politics.

Employers, too, appear keen on science education containing a portfolio of skills rather than the in-depth study of specialized topics. Mike Coles, a research fellow at the University of London's Institute of Education, has reviewed and carried out research on UK employers' requirements in science, engineering and technology education. In an address to the ASE meeting, he said that, with the exception of mathematics and physics, "employers and tutors in higher education want practical capability and application of ideas to problems to dominate qualifications, rather than simple recall".

One research group at Sheffield Hallam University is already beginning to put some of these ideas into practice. The Pupil Researcher Initiative, run by the Centre for Science Education, aims to provide pupils with an insight into current research through a series of structured activities.

Under a related scheme, 500 PhD students, known as 'researchers in residence', are sent to schools. The three-year, £1.1-million (US\$1.8-million) project is funded by the Engineering and Physical Sciences Research Council and the Particle Physics and Astronomy Research Council. It has recently been extended by two further years.

Five-and-a-half thousand UK secondary schools have been invited to join. In addition to carrying out experiments, pupils learn to write research proposals, interview scientists and contribute to journals. The objective, according to the project's deputy director Ken Mannion, is to show that science is not limited to standing behind a laboratory bench making measurements. "We try to show that science is complex, and involves the use of communication, time management and problem-solving skills."

The 'researchers in residence', according to the project's development officer, Marilyn Brodie, are encouraged to show pupils that "scientists do not have instant answers; that they need to look things up or, where necessary, acknowledge that there is no consensus".

Ehsan Masood

...as teachers try hand at research

Birmingham, England. The recent introduction of compulsory science teaching into British primary schools has had mixed impact on teachers. Although those with science backgrounds have had little difficulty in taking on the extra challenge, those lacking this background have often struggled to cope with unfamiliar ideas.

Some of the latter are now benefiting from a scheme run by Homerton College in Cambridge, and funded by the Gatsby Foundation, set up by the Sainsbury family, in which they are encouraged to undertake original research.

The aim of the 'Primary Teacher as Scientist' project, according to its director, Michael Reiss, is to "foster high-quality science education by raising the standard of primary school teachers' understanding of scientific processes".

Teachers select an area of interest and work on a research project with assistance from specialist researchers. Their findings are written up, and may even be published. One hundred teachers have signed up for projects such as why boys find overarm throwing easier than girls, and whether children suffer from seasonal affective disorder.

The teachers vary from those with no post-secondary science education to holders of postgraduate qualifications who never had the chance to pursue research. Jean McCallum, a teacher from Solihull in the West Midlands, says she became involved in order to help pass to children the experience of "doing real science". McCallum says the project has raised the profile of science among her pupils, as well as her own standing in the classroom.

E. M.

White House backs US contribution to LHC

Washington. The Clinton administration has decided to back a US contribution of \$530 million towards construction of the Large Hadron Collider (LHC) at CERN, the European particle physics laboratory in Geneva, Switzerland.

Details of the proposed contribution will be spelt out in the federal budget for the 1998 financial year (which starts on 1 October 1997), due to be unveiled next month. Administration officials will then seek congressional support for the contribution, which would be the largest ever by the United States to a scientific project overseas.

Officials at the Department of Energy have been negotiating with CERN while battling to secure US backing for the project. They received confirmation that the administration will support them early last month, when the White House Office of Management and Budget (OMB) accepted their proposed budget for high-energy and nuclear physics for 1998.

"Our challenge has been to convince OMB that our programmes deserve to be treated like those of the National Institutes of Health and the National Science Foundation," says Martha Krebs, director of research at the energy department. The Clinton administration has exempted those agencies from budget cuts. "The early indications are that we have been quite successful, at least for most of our programmes."

Krebs declines to discuss specific figures before next month's budget

announcement. But other sources say that OMB has agreed to hold steady spending on nuclear and high-energy physics at close to the 1997 level of \$1 billion.

The administration had previously published projections that would have sharply reduced spending in 1998 and later years, in line with large overall reductions that it has promised for the energy department.

The department plans to spend \$200 million on the construction of the LHC and \$250 million on its two detectors, Atlas and CMS (Compact Muon Solenoid). The National Science Foundation intends to spend an additional \$80 million or so on the detectors, although it has yet to consider its contribution in detail, and officials say it could go higher.

In a bid to secure long-term congressional commitment to the LHC, Clinton's 1998 budget will spell out the planned US contribution in full. Most of the money will be spent between 1999 and 2002, once existing construction projects have been completed at Stanford Linear Accelerator Center in California and at Fermilab in Illinois.

But each year's contribution will be included in the main US high-energy

physics budget, and will not appear as a separate line item in the budget. "We think of it as an integral part of our high-energy physics programme," says Krebs.

The US decision means that CERN will now be able to aim to complete the collider in 2005 — even though it has recently agreed to a reduction in the contributions of its member states (see *Nature* 385, 4; 1997). Without contributions from the United States and Japan, the LHC would have had to be built in two stages, with completion delayed until 2008.

The United States has not yet signed its agreement with CERN to participate in the project, which has been under negotiation since the beginning of last year. But US officials say there are no major issues to be resolved and that the agreement — which will enable the United States to participate without becoming a member of CERN — will be ready shortly.

If the collaboration goes ahead, as looks increasingly likely, it will be viewed as a considerable triumph for the US high-energy physics community in the aftermath of the cancellation in 1994 of the Superconducting Super Collider.

The last hurdle to be crossed is the Congress. Here, most key players are in favour of US participation in the collider project, and serious opposition has yet to surface. According to one supporter of the project, the only event that could derail it would be a serious outbreak of dissent within the physics community itself.

Colin Macilwain



Krebs: LHC funds 'to be part of US programme'.

Science bodies seek 7.1 per cent increase for NSF

Washington. Two leading US scientific societies are demanding an increase of 7.1 per cent in the budget of the National Science Foundation next year, arguing that this is the minimum needed to restore the spending power of its research grants to a little above their 1995 level.

The Federation of American Scientists for Experimental Biology (FASEB), whose member societies represent 43,000 biologists, backed the target at a conference last month, and it has since been endorsed by the American Chemical Society (ACS), representing 152,000 chemists and engineers.

But the American Physical Society (APS) has not yet endorsed the target, which some of its officials consider too high and too specific. Senior APS officials have written to President Bill Clinton asking for increases of 5 to 7 per cent in the basic research budgets of the Department of Energy, Department of Defense and NSF.

The Clinton administration is expected to propose lower increases for science agencies within its budget proposal for the 1998

financial year (which starts on 1 October 1997), due to be released in early February. In its initial proposal at the beginning of December, the administration's Office of Management and Budget offered flat funding to NSF and science accounts in the energy department, and a 2 per cent cut at the National Institutes of Health (NIH). After appeals from the agencies, it is now ready to propose increases of around 3 per cent for NSF and 4 per cent for NIH.

But FASEB argues that the NSF's \$3.3-billion budget for 1997 has 6.1 per cent less purchasing power than the 1995 funding because of inflation. The federation suggests that the 1998 budget should make up the gap and add an additional increment of 1 per cent. The NSF funds most non-biomedical university research in the United States.

"The NSF has taken a real whack, getting nothing for a couple of years," says John Suttie, FASEB president and professor of biochemistry and nutrition at the University of Wisconsin at Madison. He hopes the APS

and others will rally behind the target soon: "It is helpful if everyone uses the same number," he says.

FASEB had earlier announced its target of a 6.5 per cent budget increase for the NIH. The targets are intended primarily to influence appropriations committees in the Congress, which will spend the summer considering the budgets for all federal agencies proposed next month by the Clinton administration.

One appropriations staff member said the societies were deluded in expecting that NSF could do as well as NIH, which benefits greatly from pro-health-care sentiment. But he backs the view that the NSF needs stronger and better organized support from the scientific community. NSF officials agree, and privately welcome the FASEB target as a rallying point for the agency's friends in the Congress.

Later this month, FASEB, APS and ACS will publicly unveil a plan to lobby jointly for the NSF and other basic science programmes (see *Nature* 384, 393; 1996). **C. M.**

Japan's science spending climbs again

Tokyo. Despite facing the tightest budget in almost a decade, the Japanese government has agreed to a substantial increase in its spending on science and technology for the fiscal year 1997, beginning on 1 April.

Overall, science and technology expenditure across all ministries will grow by 11.9 per cent, according to figures released by the Ministry of Finance. This keeps the country on target to meet a promise made last June to boost spending on research by 50 per cent over the next five years.

The life sciences have done particularly well, with the Science and Technology Agency (STA) seeing its budgets for research in this area grow by 22.9 per cent. The Ministry of Education, Science, Sports and Culture, which now accounts for 42.9 per cent of all government spending on science and technology, will have an additional ¥10.4 billion (US\$89 million) of research grants to allocate in the year ahead.

In contrast, total government spending will rise by only 3 per cent to ¥77,390 billion — the smallest increase for nine years. The figures were approved by the cabinet last month but still have to be formally accepted by the Diet, Japan's parliament. Public finances are under severe pressure because of economic recession, the need to bail out some banks, as well as housing and loan corporations, and the rapidly growing number of elderly people in the population.

The government is determined to reduce borrowing by raising taxes and controlling public expenditure. But, despite the austere financial climate, all Japan's science-related ministries have managed to win increases in their budgets of between 4 and 12 per cent.

The increases follow the launch of Japan's new five-year plan for science and

technology, announced last June. Under the plan, the government has agreed to increase spending on science and technology to ¥17,000 billion over the next five years. Despite deteriorating government finances, the Ministry of Finance is backing the plan. "Support for scientific research is essential for Japan's economic future," says one ministry official.

The ambitious plan originated following lobbying by a small group of comparatively young and powerful politicians from various political parties (see *Nature* 378, 227; 1995). Its acceptance by the government represents agreement that such increases are needed to build new creative industries and improve the quality of life.

Several ambitious new projects in the life sciences have won financial backing. The STA will set up a brain science research institute at the Institute of Physical and Chemical Research (RIKEN) in October (see *Nature* 383, 7; 1996), in addition to receiving ¥1.9 billion for protein structure analysis, which received only ¥0.2 million this year.

The research and development budget for new industries at Japan's powerful Ministry of Trade and Industry (MITI) has more than doubled and includes ¥7.1 billion for special funding for national research institutes (up 108 per cent) of which ¥2.2 billion will be allo-

cated by competitive tender under a new initiative. MITI also gets a new budget for interministry research projects conducted at national research institutes. These funds will go into genetics, brain science, earthquake and biotechnology research. There are substantially more funds available at many ministries for postdoctoral fellowships and other schemes to encourage young researchers.

Richard Nathan

Highlights of Japan's budget for S&T 1997 (in billion yen: US\$1 = ¥114)

Science and Technology Agency (STA)

Total budget	734.5	+6.0%
General R&D budget	571.4	+8.0%
Space	180.6	+1.5%
Spring-8	19.2	+15.4%
Human genome	2.9	+17.5%
Protein structure analysis	1.9	+688.0%
STA fellowships	3.1	+23.3%
Brain	9.9	+296.0%

Ministry of Trade and Industry (MITI)

Total R&D budget	421.3	+12.1%
Agency for Industrial Science & Technology	162.0	+13.8%
New industry R&D	26.8	+113.0%
Industrial scientific technology	28.0	+6.2%
Joint ministry research projects	3.4	New
Human Frontier Programme*	4.0	+12.0%

* includes 2.4 billion yen from STA

Ministry of Education, Science, Sports and Culture

Grants in Aid for Scientific research	112.2	+10.2%
Japan Society for Promotion of Science	39.0	+48.2%
(postdoctoral fellowships)	(14.4)	(28.0%)
Joint research with industry	48.1	+37.8%
Academic information network	36.2	+9.5%
Centre of excellence (COE) programme	13.4	+18.3%

France says goodbye to the fast-breeder as Superphénix

Paris. Almost four decades of effort by France to build fast-breeder reactors to produce electricity commercially from plutonium came to an end on Christmas eve, when its 1,250-MW Superphénix prototype power plant in Creys-Malville was shut down. The plant will be converted into a research reactor intended to burn up plutonium from waste and dismantled weapons, as well as actinide wastes (see *Nature* 365, 381; 1993).

Ironically, although 1996 was the last year in which Superphénix operated as a commercial prototype, it was the first year in which the troubled reactor operated smoothly. The 3 billion kilowatt-hours of electricity it generated equalled the total produced from its entry into service in 1986 until the end of 1995 — over this period, technical problems meant that it operated for a total of only ten months.

Now, 72 of the 220 fuel elements in the core will be replaced with stainless-steel dummies to reduce the reactor's capacity to produce plutonium. At the same time, three experimental fuel elements will be placed in the core in preparation for future research. Two will contain plutonium configured for experiments to boost the reactor's capacity to incinerate plutonium. The other will contain 2 kg of the highly radioactive actinide, neptunium. The reactor will begin to incinerate plutonium only around 2003, when it will be equipped with a core lacking a uranium blanket.

The conversion marks the end of the last programme in Europe to develop a generation of fast breeders for electricity production. A project to build a European Fast Reactor has also been abandoned following the withdrawal of the United Kingdom and Germany in 1992 (see *Nature*

360, 93 & 703; 1992). Only Japan is now actively pursuing the technology (see *Nature* 379, 196; 1996).

But the French government's decision in 1994 to reincarnate Superphénix as a research reactor has not dampened the controversy surrounding it. At the time, the decision was widely viewed as a political solution to the need to satisfy France's partners in NERSA, the consortium of French, German and Italian utilities that built Superphénix in the 1970s at a cost of FFfr27.7 billion (US\$5.2 billion) (see *Nature* 368, 281; 1994).

The French government has since agreed to pay its partners in the project an estimated FFfr1 billion annually — in the form of electricity — to compensate for loss of income following the conversion. To these costs must be added that of the planned research programme (FFfr500

'Misconduct' dispute raises fears of litigation

San Diego. A legal dispute with potentially far-reaching implications for the way universities investigate scientific misconduct cases has been generated by the conclusion of a report from Baylor College of Medicine in Houston, Texas, that one of its researchers published data that had been fabricated.

The dispute has raised the question of whether universities should be immune from certain types of civil lawsuit arising from investigations of research misconduct carried out as required by National Institutes of Health (NIH) regulations.

The dispute centres on Kimon J. Angelides, a neurobiologist who was fired by Baylor in early 1995 after officials concluded that he had falsified research results described in five published papers and in documents used in applications for NIH grants (see *Nature* 383, 107; 1996).

Angelides subsequently moved to the University of Durham in England. He filed a civil lawsuit against Baylor, the scientists who had accused him of improprieties, and members of various *ad hoc* committees at Baylor that ruled that he had committed scientific misconduct.

The civil lawsuit is for wrongful termination, breach of contract, defamation and other actions. These include an allegation of 'blacklisting', a charge associated with the fact that Baylor reported Angelides' alleged offences to the NIH's Office of Research Integrity (ORI), the body that investigates allegations of research misconduct made against scientists receiving NIH funds.

This lawsuit is before a federal appeals court, where Angelides, Baylor and other interested parties are arguing about whether a university should be immune from certain civil lawsuits when investigating misconduct.

takes on new role

million annually), the reactor's annual running costs (around FF£1 billion) and the FF£1.25-billion cost of a new core. The French national audit commission recently estimated that the reactor will have cost a total of FF£60 billion by 2000.

But continuing to operate the reactor is considered to be cheaper than shutting it down. A shutdown would entail the government paying as much as FF£18 billion in compensation to NERSA as a lump sum, as well as the costs of decommissioning.

Whatever the outcome of the research, few expect France to invest in the short term in the many fast breeders that would be needed in practice to eliminate the large stocks of plutonium. Superphénix, which will eventually shut down for good around 2015, seems likely to be one of the last of a dying breed.

Declan Butler

In addition to the civil litigation, federal prosecutors in Houston have opened a criminal inquiry into Angelides' research activities. This raises the possibility that he could be indicted for defrauding the NIH of research funds.

Such criminal investigations are rare, with only one scientist in recent years having been convicted of a crime related to research misconduct.

In the civil case, the federal requirement that a university that receives NIH grants should monitor its scientists, investigate allegations of misconduct and report the conclusions is clashing with an accused scientist's right to due process.

Angelides denies any research misconduct, blaming others for any discrepancies in his reported results. His attorneys argue that he should be able to sue his accusers for damages in a Texas state court, where his civil lawsuit was filed.

But Baylor's attorneys reply that the university is immune from such state court civil action as it is acting for a federal agency, the NIH. They also argue that if the verdict of a civil lawsuit is to be based on the propriety of a university probe of a scientist, the case should be heard in a federal district court, not a state court.

Such issues have already attracted much interest from the Association of American Medical Colleges (AAMC), as well as the ORI, both of which have filed briefs supporting Baylor's position.

Chris B. Pascal, acting director of the ORI, says his agency fears that if Angelides' method of suing is upheld, an already difficult process — the scientific review of misconduct charges at the university level — will be significantly damaged, if not destroyed.

Angelides' lawsuit against Baylor threatens the "partnership" between the NIH and research institutions, says Pascal. "This is a big issue for us," he adds. "If an institution and its committees of scientists are held liable for reporting to ORI, who would write a report and put their name on it? They wouldn't do it any more."

Similarly Joseph A. Keyes Jr, general counsel to the AAMC, says his members believe that state court lawsuits filed by scientists whose research has been challenged by their peers would have a chilling effect on the ability of universities to fulfil their obligations to funding agencies to monitor the conduct of research.

"Faculty members and [other] individuals called to serve on investigating committees fear being tied up in endless litigation," says Keyes, whose organization represents 400 universities and medical schools and 90 professional societies.

Such arguments are being made in legal briefs before the US Court of Appeals for

the Fifth Circuit in New Orleans. A decision on the immunity issue is expected in a few months, and could have broad implications for scientific misconduct cases elsewhere in the United States.

Meanwhile, the trial in Angelides' civil lawsuit has been postponed until August, largely because of the criminal probe of his activities. The existence of the criminal inquiry surfaced last month during a hearing on the civil lawsuit.

Angelides had flown from the United Kingdom to Houston to be questioned by Baylor's attorneys in a deposition. But,



during a brief hearing with the attorneys, Angelides' attorney argued that the pending criminal probe might use his sworn deposition testimony against him, and invoked his Fifth Amendment privilege against self-incrimination.

David H. Peck, the assistant US attorney in Houston who is conducting the inquiry into Angelides' actions, declines to comment on the federal probe. Court records indicate that federal authorities are expected to make a decision on whether to file criminal charges within a month or so.

Angelides declines to comment on his situation. But Rusty Hardin, his criminal attorney in Houston, argues that, for a criminal violation, "someone must conclude that he [Angelides] knowingly submitted false information" to the NIH, which he "unequivocally" did not.

"There is no issue [that] some wrong data was submitted," says Hardin. "The question is: did he know it was wrong when he submitted it? I am satisfied there is no criminal violation."

Pascal at ORI declines to comment on how the criminal inquiry may affect his agency's investigation of Angelides, which started in spring 1995. The ORI cooperates with federal prosecutors when they conduct criminal inquiries, he says, and may delay or complete its investigation, depending on what the federal prosecutors may want.

Rex Dalton

Biodiversity projects face funding challenge

London. The principal United Nations agency responsible for funding environmental programmes — including environmental research — has had to reduce its support for biodiversity projects partly because of a lack of high-quality applications, particularly from developing countries.

Last year, the agency, known as the Global Environment Facility (GEF), awarded \$23 million for biodiversity projects, \$42 million less than in 1995. During the previous four years, in comparison, biodiversity activities were awarded on average \$80 million each year by the GEF.

Both government representatives responsible for overseeing the UN biodiversity convention and scientists working in the area have begun to raise the issue. According to one government delegate to the biodiversity convention, concern is also mounting among members of the GEF's governing council. "There is enough money," he says. "There are just not enough quality proposals."

Stephen Blackmore, keeper of botany at the Natural History Museum in London, says that developing countries are caught in a bind. The GEF's guidelines imply that projects should be initiated by countries themselves. But this handicaps developing countries, whose scientists may have diffi-

culty in preparing acceptable research proposals. Blackmore believes projects should also be allowed if proposed jointly with an institution from a developed country.

Simon Owens, keeper of the herbarium at the Royal Botanic Gardens at Kew in London, agrees, and says there is considerable willingness on the part of research organizations, such as his own, to contribute to technology transfer. "We are not into being colonial any more," he says.

Mario A. Ramos, a senior environmental specialist with the GEF in Washington DC, accepts that the reduction in take-up of biodiversity funds can be partly blamed on the lack of expertise in developing countries to prepare effective projects. But he says that he expects this situation to change as more countries complete their 'biodiversity action plans', in which signatories to the convention identify national priorities in biodiversity conservation and research.

Ramos also points out that there is no bar to nongovernmental organizations jointly proposing projects with countries, and that several — including Kew Gardens — are already involved with technical management and assistance. He adds that many factors help to account for the shortfall in biodiversity proposals. "Quality projects take

time to prepare," he says. "One of the GEF's requirements is a high degree of public participation in the planning stages. It could take up to two years for funding approval after you have solicited the views of nongovernmental organizations, local communities and the government."

The GEF was set up in 1991 to provide grants and concessional funds for projects that aim to protect the environment. The facility has 156 members, and spent \$733 million on 115 projects during its pilot phase between 1991 and 1994. In March 1994, 34 members pledged to replenish the fund with an additional \$2 billion.

GEF funds are available only to countries eligible for loans from the World Bank group of financing agencies, or technical assistance grants from the United Nations Development Programme (UNDP). In addition to climate change and biodiversity, a smaller proportion of GEF funds are also spent on projects related to international waters (12 per cent) and protection of the ozone layer (10 per cent).

GEF spending is administered by the UNDP, the UN Environment Programme (UNEP) and the World Bank. The UNDP looks after projects that involve technical assistance, while UNEP is responsible for

Does yur autoated
DNA seqencer leave
u guessing?

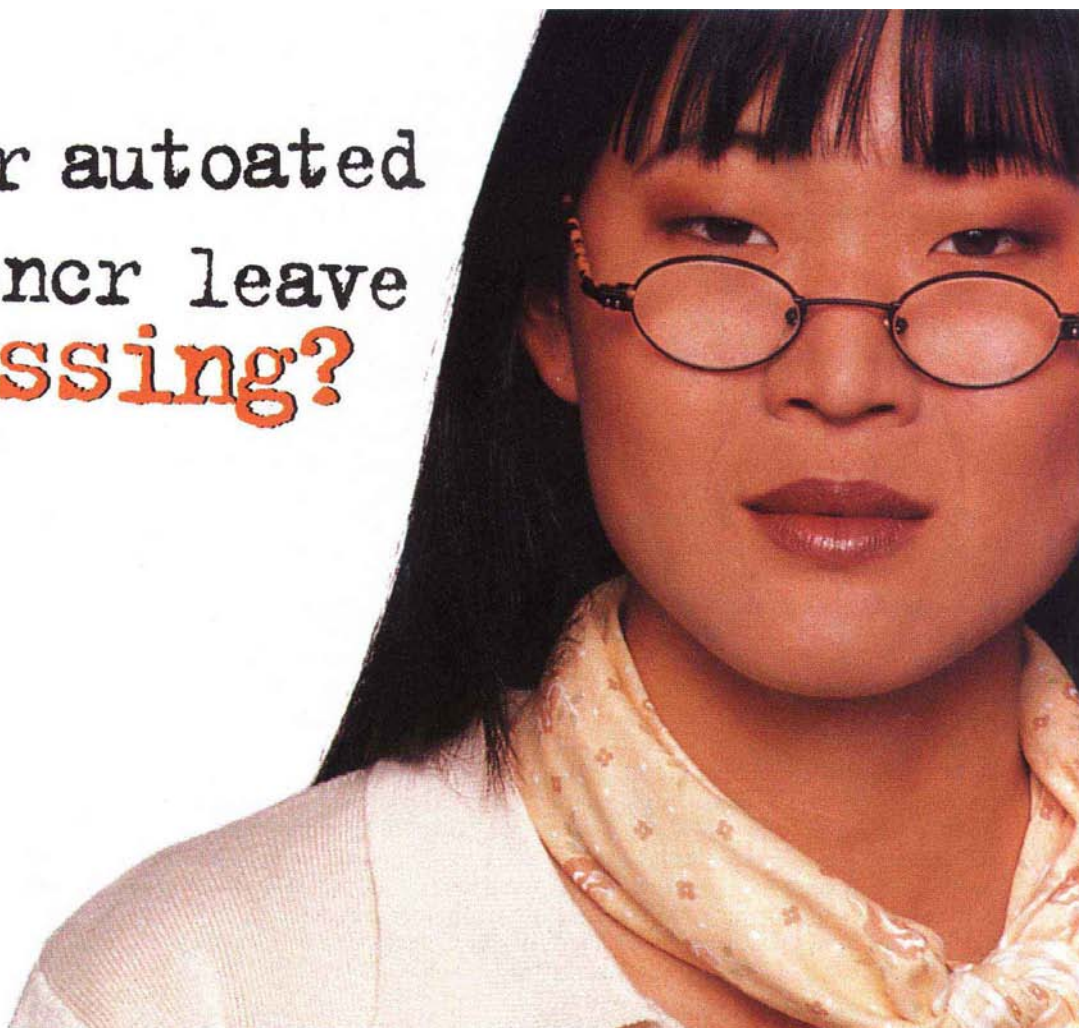


IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Sustainable agriculture: but more research is needed.

scientific and technical analysis. The World Bank, however, remains the senior partner.

The bank is responsible for 'investment' projects, such as managing national parks or developing non-fossil fuel energy sources, which constitute 70 per cent of GEF spending. The GEF secretariat is based at the bank's headquarters in Washington.

In the past, tensions have arisen between the GEF and members of the biodiversity convention, particularly over final authority to approve project finance. The biodiversity convention states that its parties will determine policy, strategy, programme priorities and eligibility criteria for biodiversity finance. This is interpreted by developing countries to suggest that the parties should be the final arbiters of project allocation.

But GEF council members from the

developed world disagree and have resisted what they consider to be attempts at 'micromanagement'. "You cannot have 160 countries all deciding on which projects to fund," says one delegate from a developed country.

A 'memorandum of understanding' between the two sides was eventually brokered at the annual conference of the parties to the convention in Argentina last November. The memorandum acknowledges that the convention is responsible for policy and strat-

egy. But the GEF will decide on project selection and management within agreed guidelines.

Contentious areas remain. Biodiversity environmentalist groups, some scientists and some developing countries are still not satisfied with the GEF's decision to increase its climate change budget at the expense of biodiversity. Owens of Kew Gardens says that "biodiversity is sometimes considered the poor relation to the more fashionable climate change", but is just as important.

During 1991-94, the GEF spent 45 per cent of its budget on biodiversity projects and 35 per cent on projects in areas that reduce the impact of climate change, such as power generation. During 1995 and 1996, these percentages swung to 19 and 47 per cent respectively. The GEF now plans to

devote around 30 per cent of its budget to biodiversity, and 45 per cent to projects related to climate change for 1997 and 1998.

Ramos says criticism about the apparent imbalance is misplaced. Climate change projects, such as renewable energy sources, tend to be more capital-intensive than biodiversity projects, he says.

So far, he adds, GEF has provided \$415 million in biodiversity project finance for more than 71 national, regional and global projects. But, he says, it expects a greater number of biodiversity projects during the next few years — more than climate change — and is setting aside \$136 million for such projects during 1996-97.

Another controversial issue is the nature of the GEF's financing method. GEF does not fund projects in full, but only the difference — the 'incremental cost' — between the price of a project, such as a coal-fired power station, and the cost of an environmentally cleaner alternative.

Blackmore of the Natural History Museum says incremental funding is inappropriate for research proposals and feasibility studies. But Ramos counters by pointing out that GEF operates a separate fund for project analysis and feasibility studies that does not operate on the incremental cost model. Between \$25,000 and \$1 million is available to assess the feasibility of any proposal, including the conduct of relevant research.

Ehsan Masood

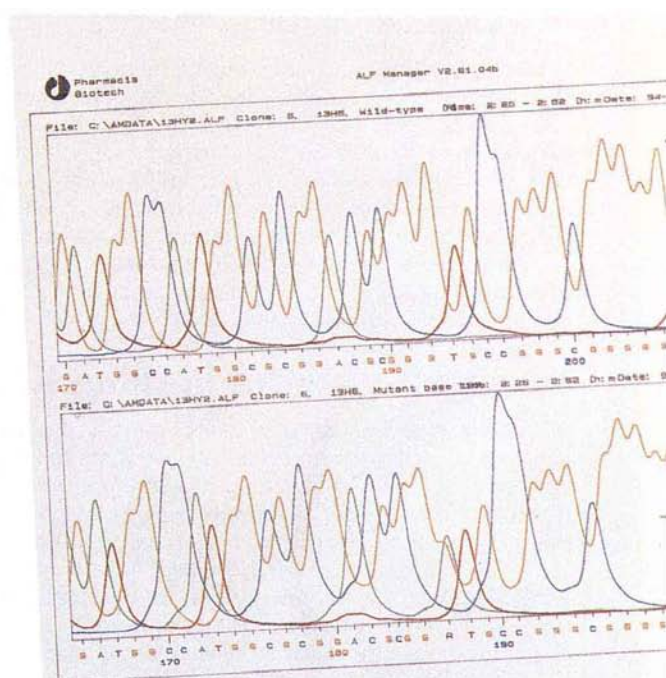
Does your automated DNA sequencer leave you guessing? If so, chances are it's primarily designed for high throughput sequencing. Why be uncertain of your sequencer's accuracy, when ALFexpress™ is providing researchers with the full genetic stories of their DNA.

ALFexpress: for more accurate readings

In the largest clinical study using automated DNA sequencing, the technology behind ALFexpress proved exceptional (see caption). That's one of many examples of ALFexpress offering unrivaled accuracy during automated confirmatory sequencing. Further, its readings are so accurate that ALFexpress can unambiguously identify heterozygous point mutations—as proven in many clinical research applications, such as analysis of tumor genes and high-resolution HLA typing. What's more, Pharmacia Biotech has dedicated software programs to support these applications.

For the full story, call us: 1 (800) 526-3593 from the U.S.; +81 (0)3 3492 6949 from Japan; or +46 (0)18 16 50 00 from Europe and the rest of the world. Or visit us on the Internet: <http://www.biotech.pharmacia.se>.

**Pharmacia
Biotech**
Uppsala, Sweden. (And the rest of the world)



The p53 gene from 316 breast cancer patients was sequenced using ALF automated sequencing technology. (Bergh J., Norberg, T., Sjögren, S., Lindgren A., Holmberg, L. "Complete Sequencing of the p53 Gene..." *Nature Medicine* 1995; 10:1029-1034.)



US approval expected for space station control module

Washington. US space officials expect approval this month to build an 'Interim Control Module' (ICM) that would keep construction of the international space station on track despite delays with a crucial Russian module. US space station managers were told last month that the Russian Service Module, which will maintain the station in its proper orbit during assembly, will be at least eight months behind its scheduled April 1998 launch. Construction of the station begins in November.

The less capable ICM, which is being designed by the US Naval Research Laboratory based on an existing defence satellite 'bus', would be paid for out of reserve funds, says Wilbur Trafton, head of space station planning at the US space agency. It would therefore not break the project's strict \$2.1-billion annual spending cap. But the plan still needs approval from the White House and Congress. □

Gene therapy struggles in Europe

London. World-class work by Europe's gene-therapy researchers is not being translated into commercial applications because biotechnology companies are not being set up to exploit these innovations, according to a report from the Science Policy Research Unit at the University of Sussex.

The report, prepared for the European Commission by Sandy Thomas and Paul Martin, contrasts the small number of gene-therapy companies in Europe (12) with the 40 to 50 such companies in the United States. The authors point out that European pharmaceutical companies are investing in the US market and creating skilled jobs there, with the result that the benefits of European research are being lost to European companies. □

Taxonomic alliance formed

London. Ten European taxonomic institutions have formed a consortium to help realize the joint potential of their natural history collections and expertise. The Consortium of European Taxonomic Facilities aims to increase the efficiency of taxonomic facilities through networking and cooperation, and its objectives will include improving the scientific, commercial and public access to databases, and promoting the training of systematists. Founding members of the consortium are museums in Copenhagen, Frankfurt, Amsterdam, Paris, Madrid, Berlin, Stockholm and London and the Royal Botanic Gardens at Kew, also in London. ▲

Transgenic centre in Hong Kong

Hong Kong. A regional centre to train scientists in the production of transgenic animals has been set up at Hong Kong University, in the hope of boosting the study of genetics in the whole Asia-Pacific region. The university's biochemistry department believes it is the first in the region to have carried out gene-targeting work.

"We are trying to make this technology more available to scientists within Asia," says Kathryn Cheah, co-organizer of the centre and reader in biochemistry. "Our long-term aim is to make Hong Kong the centre of transgenic technology for the region." □

French agency leaves Mururoa

Paris. The French Commissariat à l'Énergie Atomique (CEA) closed a 34-year chapter in its history on 31 December when it withdrew from Mururoa Atoll in the South Pacific, France's nuclear weapons testing site. Its departure follows the decision last year by Jacques Chirac, the French president, to abandon explosive nuclear testing and to sign the Comprehensive Test-ban Treaty.

CEA will retain a virtual presence at Mururoa, however, having left behind an automatic surveillance network, Telsite, which will

relay data to France via satellite from a dozen stations around the atoll. Telsite will provide geophysical data, as well as estimates of radioactivity in the air and water. Annual visits will also be made to take samples from the environment and from plants and animals.

Meanwhile, the International Atomic Energy Agency is nearing completion of an independent study of the consequences of testing in the region, and is due to release its report at the end of this year.

Dismantling of the CEA base at Mururoa — the Centre for Pacific Tests, which opened in 1962 and supervised 193 nuclear tests — will be completed by mid-1998, at an estimated cost of FF135 million (US\$25.6 million). France will also pay around FF1 billion annually over ten years to French Polynesia to compensate for closure of the test site. □

Mary Leakey fund established

London. A fund has been set up in recognition of the scientific contributions of the palaeoanthropologist Mary Leakey, who died last month, to provide support for scientists pursuing interests similar to hers in fields such as archaeology, anthropology and palaeoanthropology. The creation of the Mary Leakey Fund for African Archaeology has been announced by the board of trustees and scientists of the L. S. B. Leakey Foundation, which was itself set up in 1968 in honour of her late husband. □

US ban to protect against BSE

Washington. The US government has proposed a wide-ranging prohibition of the use of meat in animal feed, as a precaution against an outbreak of bovine spongiform encephalopathy (BSE), or mad cow disease, in the United States. The ban would prohibit the use of tissue from cows, sheep or goats in feed but permit the continued use of blood, milk and gelatine, for which no infection pathway is known, in feed for these animals.

If approved after a consultation period, the ban will codify a voluntary ban already being observed by the meat industry. Donna Shalala, the health secretary, said last week that, although there have been no reported cases of BSE in the United States, the ban "would add another level of safeguards to protect the United States against the potential risk". □

Walker takes lobbying post

Washington. Robert Walker, the former Republican chairman of the Science Committee in the US House of Representatives, has been appointed president of Wexler Group, a Washington lobbying firm. Walker was a senior figure in the House Republican leadership until his surprise announcement a year ago that he would retire after the 104th Congress, which has just ended (see *Nature* 378, 757; 1995). He began work at Wexler this week, as the 105th Congress convened. Wexler's clients include the Science Coalition, a pro-science lobbying operation set up by some universities, and industrial corporations ranging from General Motors to Burger King. □

Russian records rise in HIV

Moscow. Russia registered 1,031 new HIV carriers during 1996, almost equal to the total number of carriers for the previous nine years, according to a report from the Interfax news agency. Most of the new cases contracted the virus through the sharing of syringes for drug use. Russia now has 2,316 HIV carriers registered since 1987. The report quoted a health ministry official warning that a lack of funds was obstructing efforts to increase public awareness of AIDS. The number of patients in the "last and incurable stage of the disease" stands at 248.

Meanwhile, Hungary's National Health Institute recorded 606 new HIV carriers during 1996, according to a report from the MTI news agency in Budapest. Two hundred and thirty four have since developed AIDS, bringing the total number of AIDS patients to 4,000. Hungary's AIDS deaths now total 151. □

Safety of transgenic corn

SIR — We feel that a few clarifications are needed on the main topic of a recent leading article and the News story by Meredith Wadman (*Nature* **383**, 559 & 564; 1996).

(1) Ciba's transgenic corn is first and foremost an insect-control tool provided to farmers so they can protect their crops against one of the most devastating corn pests, the European corn-borer. Both articles emphasized the herbicide-resistance trait as the main feature of Ciba's transgenic corn (maize).

(2) Britain's Advisory Committee on Novel Foods and Processes was concerned only about the remote possibility of transfer of the nonfunctional β -lactamase DNA sequence from Ciba transgenic corn to gut microbes of livestock to which it is fed. Some of your comments extended this concern to the possibility of transfer to humans, thereby presenting an inaccurate view of this issue.

(3) The risk of transfer of β -lactamase DNA from plant material to gut microflora of livestock is vanishingly small, and is far outweighed by the economic, environmental and feed safety benefits of this product. The only country that has so far formally raised the transfer of ampicillin resistance as an issue is the United Kingdom. Ciba's *Bt* corn has been reviewed and fully approved by the US, Canadian and Japanese governments. In addition, the French authorities who acted as rapporteur for our dossier in Europe also recommended full commercial approval. Genuine scientific debate is taking place, and outside the United Kingdom there is a scientific consensus emerging on the topic which was not echoed in your leading article. Several prominent independent scientists have expressed their views on the subject. Furthermore, the Foundation for Nutritional Advancement and Tufts University as well as a joint FAO/WHO Expert Consultation on Biotechnology and Food Safety concluded that the issue of ampicillin-resistance transfer from a plant to a microbe constituted a near-zero risk of creating ampicillin-resistance complications. Antibiotic-resistant microbes are much more likely to occur by DNA transfer from one bacterium to another in nature.

(4) The "aggressiveness" that you correctly ascribed to Ciba reflects our strong affirmation of the scientific consensus on the safety of our *Bt* corn. Contrary to the impression left by your leading article that Ciba has remained insensitive to concerns raised by advisory and regulatory committees, Ciba's scientific and regulatory staff have worked cooperatively at every opportunity with the appropriate committees in matters concerning the benefits and safety of genetically modified crops worldwide. In this regard, Ciba's approvals for production and commercialization of its transgenic corn

in the countries mentioned have been obtained through a cooperative effort using sound scientific data and rationale to address the various issues and concerns.

Juan J. Estruch
Mary-Dell Chilton
Richard Lotstein*

Ciba Biotechnology,
3054 Cornwallis Road,
Research Triangle Park,
North Carolina 27709-2257, USA

Wallace Beversdorf
Ciba Ltd, Seed Division,
Basel, Switzerland

*To whom correspondence should be addressed.

Live universes

SIR — In their recent Commentary article (*Nature* **384**, 107; 1996), John Maynard Smith and Eors Szathmari consider the idea (proposed by L. Smolin and J. A. Wheeler) that universes are 'alive' in the sense that they can pass on hereditary information, show variability from one generation to the next, and are subject to selective pressures. It seems ironic that evolutionary biologists should take such an idea seriously enough to comment (even sceptically) on it, while at the same time treating with derision the much less radical proposals that ecosystems or planets (in particular, Earth) may also be alive.

The possibility that life forms exist at levels above that of individual organisms is generally denied because these higher-order systems fail one or more of the criteria established for the recognition of life. But this seems tautological; there is no reason to expect that super-organisms would meet criteria based on observations of individual organisms. Isn't it time to consider the possibility that the boundary between life and non-life may be diffuse, non-stationary over time, and dependent upon scale?

M. G. Björnerud
Geology Department,
Lawrence University,
Appleton, Wisconsin 54912, USA
e-mail: bjornerm@lawrence.edu

SIR — Maynard Smith and Szathmari say that the simplest interpretation of the strong anthropic principle is that the Universe was created purposefully. They make the intriguing claim that "this interpretation lies outside science" and turn their attention to more complex and (one presumes) more 'scientific' interpretations.

What do they mean by stating that a hypothesis is "outside science"? They offer no justification for this bald statement and make no attempt to persuade the reader that the other hypotheses under considera-

tion are in any sense better or more meaningful. Simply to dismiss a concept as being beyond analysis seems a rather crude form of intellectual censorship. Those who seek to address the possibility of God's existence in a rational way are thereby disenfranchised.

If the Universe has been purpose-built, one might expect there to be evidence of this. At the very least we should address the data without excluding this possibility. We may or may not see God's signature, but we owe it to ourselves to try to keep our eyes open.

Graeme Suthers
South Australian Clinical
Genetics Service,
Centre for Medical Genetics,
Women's & Children's Hospital,
North Adelaide,
South Australia 5006
e-mail: suthersg@wch.sa.gov.au

Out of context

SIR — In the space of one paragraph, George Magyar (*Nature* **384**, 509; 1996) accuses me of trying to discredit science and three of its more notable contemporary practitioners: Steven Weinberg, Herbert Simon and Noam Chomsky. He makes this accusation in the context of complaining about the unscholarly behaviour of sociologists of science. Unfortunately, his references to my work in *Social Studies of Science* are either completely spurious or twisted out of context.

It is true that I have had occasion to criticize Weinberg and Simon in reviews of their work (as part of an open exchange with them). I suggested, in the case of Weinberg, that to identify science exclusively with what élite scientists do is comparable to "megalomania" (*Soc. Studies Sci.* **24**, 154; 1994). In the case of Simon, when I said he had a "muddled grasp of the metatheoretical issues surrounding his work" (**21**, 149; 1991), I was simply making a point that philosophers have raised before, namely that Simon equivocates over whether his computer simulations are meant to *describe* or to *prescribe* the processes of rational thought. As for Chomsky, my reference to him as a "linguistic upstart" (**19**, 629; 1989) was not a judgement about the quality of his work but a description of how behaviourists initially viewed him when he launched his famous attack on B. F. Skinner.

I am content to let readers draw their own conclusions about my claims in the context of the original discussions.

Steve Fuller
Department of Sociology
& Social Policy,
University of Durham,
Durham DH1 3JT, UK
e-mail: steve.fuller@durham.ac.uk

Confucianism oversimplified

SIR — Your leading article “Can Confucius excuse poor creativity?” (*Nature* 384, 197; 1996) suggests that Confucian thoughts are partly to blame for lack of creativity in Korea. I believe that this oversimplifies Confucian ideas.

The original texts of *Lùn Yǔ* (Analects), *Zhōng Yōng* (The Great Mean) and *Dà Xué* (Great Learning) constitute the gist of Confucian philosophy. They are records of what Confucius said or did during his lifetime. His philosophy is not an authoritarian philosophy where new ideas are not accepted, nor did he believe in blind devotion to authority.

In *Dà Xué*, Confucius said: “The perfecting of knowledge depends on the investigating of things. If we wish to carry on knowledge to the utmost, we must investigate the principles of all things we come into contact with, for the intelligent mind of man is certainly formed to know, and there is not a simple thing in which its principles do not inhere. But so long as all principles are not investigated, man’s knowledge is incomplete.” (James Legge’s translation, for this and for other quotations.)

And in *Lùn Yǔ*, it was recorded that “there are four things from which the Master was entirely free. He had no foregone conclusions, no arbitrary predetermination, no obstinacy and no egoism.” This is not a scholar unwilling to accept new ideas; on the contrary, his ideas about investigation and knowledge are almost identical to what we would call ‘the scientific method’ today.

Nor is blind respect for authority part of what Confucius advocated. He allowed his students to have their way, even when this was against the traditions of their times. *Zhōng Yōng* recorded him as saying that “sincerity is the end and beginning of things. Without sincerity there would be nothing.” And when it pertains to knowledge, *Lùn Yǔ* recorded him as saying: “[Shall I teach you what knowledge is?] When you know a thing, to hold that you know it; and when you do not know a thing, to allow that you do not know it — this is knowledge.” That is intellectual honesty.

Confucian thought, or strictly speaking, the philosophy of *Rú Jiā*, although lacking the logical rigour of many Western philosophical ideas, is a self-consistent philosophical system. Over 25 centuries, this system of philosophy has probably influenced more people than any other. The three main proponents behind it are Confucius, Mencius and Xún Zǐ. Even during that time, one could already discern slight differences between these three philosophers. The system of thought evolved during the Han Dynasty (206 BC to AD 220) to be “polluted” by astrological ideas. During the past 2,000 years or so, the philosophy of *Rú Jiā*, either in territories inhabited by the Chinese or elsewhere, has been subject

to reinterpretation, misinterpretation, mutilation, disfigurement and occasionally enrichment. The philosophy of *Rú Jiā* spread to Japan, Korea, Vietnam and other Asian countries, but each country adapted it to suit its own cultural background. The Confucian philosophy in Japan or Korea today is a far cry from the original ideas of this native of the Duchy of Lǔ.

If people read only *Lùn Yǔ* and try to understand Confucius’s original ideas, they will see that he was a gentle, learned man, filled with curiosity about the world around him, never failing to see beauty in nature, understanding the youth of his day, and always emphasizing truth and sincerity. The outward show of things was never as important to him as the inward feelings. He was not a person who would demand that his students bow to him. Although his ideas did not explicitly mention imaginative and creative thinking, he certainly did everything to encourage it.

In your last paragraph, you ask: “So how to boost scientific creativity?” Perhaps we should return to the original ideas of Confucius, to emphasize the importance of investigation in our acquisition of knowledge, to encourage intellectual honesty and to avoid the four enemies of scholarly work of any kind: foregone conclusions, arbitrary predetermination, obstinacy and egoism.

At a time when people are uncovering more and more cases of scientific misconduct, Confucian standards of sincerity and intellectual honesty would be our best guardians against “dubious science” of any kind.

P.-L. Chau

Kemisk Laboratorium III,
H. C. Ørsted Institutet,
København Universitet,
DK-2100 København Ø, Denmark
e-mail: chau@st.ki.ku.dk

W. Y. Chau

30/74 Wardell Road,
Sydney 2206,
Australia

Bias in Japanese university awards?

SIR — I welcome the report¹ of the new scheme, “Research for the Future”, which has been set up this year with an initial budget of US\$100 million from the Japanese Ministry of Education to promote university-based research in Japan. It is particularly good news that these grants can be used flexibly to employ research workers. This is an important development, because traditionally all grants (including those from the government) have been limited to the purchase of hard-

ware such as laboratory equipment, and have not been available for salaries.

It may, however, surprise readers, especially those from the West, that these project grants will be distributed in a ‘top-down’ fashion by a hierarchy of committees. Naturally, this may lead to suspicion of bias in the selection of project leaders, and this is referred to in the report. In the West, investigator-initiated peer-review has proved to be a successful system of grant allocation². Furthermore, this is the only system that allows “grant applications to be approved on the basis of scientific merit, rather than to meet an administrative or political agenda”³. Perhaps a lesson can be learned from Italy, where the administration of an AIDS research programme has successfully introduced a system of grant appraisal based on peer review by Italian and foreign scientists³.

Transparency or disclosure (*‘Kaiji’* in Japanese) of the administrative process at all stages is increasingly demanded by the public in Japan. Taxpayers demand value for money in all areas, and the allocation of scientific funding is no exception. It would therefore also be wise to have a system for monitoring the productivity of grant recipients, as do established grant-awarding organizations in Western countries².

Noriyoshi Takei

Institute of Psychiatry & King’s College
School of Medicine,
De Crespigny Park,
London SE5 8AF, UK
e-mail: spjunot@iop.bpmf.ac.uk

1. Swinbanks, D. *Nature* 383, 369 (1996).
2. *Lancet* 344, 488–489 (1994).
3. Abbott, A. *Nature* 374, 299 (1995).

Italian grants

SIR — Alison Abbott says (*Nature* 383, 567; 1996) that in Italy “many grant-giving agencies and charities distribute small amounts of money evenly, to avoid upsetting unsuccessful applicants”.

Italy invests only 1.1% of gross domestic product in scientific research, and most of this money is allocated for targeted projects, usually without peer review, to a minority of scientists. A negligible amount is available for basic research, where there is a huge number of applications. But even then the distribution is not even. In 1996, for example, the biomedical committee of the National Research Council distributed an average of US\$8,000 to only 22% of applicants. Private institutions such as Telethon use peer review by international referees. The distribution is not widespread, and it is a much more satisfactory system.

Piergiorgio Strata

Università di Torino,
Dipartimento di Neuroscienze,
C.so Raffaello 30,
I-10125 Torino, Italy

In search of molecular darwinism

Paul M. Sharp

Attempts to detect adaptive evolution at the molecular level have met with little success. Studies of a digestive enzyme in primates, involving the reconstruction of DNA sequences that have long been extinct, show a way forward.

THE molecular genetics revolution has transformed evolutionary biology, but with rather more success in some areas than in others. About 70 years ago, in an essay predicting the future of biology, J. B. S. Haldane wrote¹ that "our ideal is to establish a family tree of plants and animals", in which he included being able to state when in the past the common ancestors of extant species lived. Haldane suggested that this could be achieved by chemical methods, and he was right, even if his timescale was not: Haldane predicted that it would take thousands of years, whereas phylogenetics has already been the outstanding success story of molecular evolution. Haldane foresaw not only the charting of the evolutionary past, but also the prospect of finding evidence for the central mechanism of darwinian evolution, namely adaptive change. He looked forward to being able to ask "What inheritable variations ... show any sign of being ... of advantage to their possessor?". In this respect, molecular data have, as yet, thrown up disappointingly few examples.

One apparent success concerns the enzyme lysozyme in primates. Colobine monkeys are leaf-eaters with a fermentative foregut where lysozyme is used in bacteriolysis in much the same way as in ruminants. Some years ago, it was shown that in one colobine, the hanuman langur (Fig. 1), lysozyme has amino-acid differences from other primates that parallel changes seen in the cow². That the same amino-acid changes had occurred in two independent lineages provided firm

hints that they were provoked by parallel selective pressures.

Messier and Stewart (page 151 of this issue³) have now examined the evolution of primate lysozyme in much more detail

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

FIG. 1 Foregut in ferment — the hanuman (or common) langur (*Presbytis entellus*), one of the colobine Old World monkeys that have a foregut adapted to the digestion of leaves.

— they have pinned down the proposed adaptive changes to a short period during the early emergence of colobine monkeys (see the authors' Fig. 1 on page 152), and provide unexpected evidence for another period of adaptive evolution in the ancestors of hominoids (Fig. 2, page 153). Their approach has been to contrast the rates of non-synonymous and synonymous substi-

tutions (K_A/K_S ratios; see box) among the lysozyme genes from a wide range of primate species. Comparisons of K_A and K_S have often been used elsewhere, but for reasons that are (with hindsight) fairly obvious, have identified few cases of adaptive molecular evolution.

Adaptive changes will be very hard to find using K_A/K_S ratios, particularly if we only look for ratios greater than one (see box). Genes encoding functional proteins are always expected to be subject to purifying selection, weeding out most amino-acid changes because of their detrimental effects, and thus lowering the value of K_A to substantially less than K_S . Consequently, K_A/K_S ratios may often be less than one even if some adaptive substitutions have occurred. For example, when 363 equivalent gene sequences were compared between mouse and rat⁴, only one showed a ratio greater than one (Fig. 2, overleaf); yet it would be most surprising if this were the only one of these genes that had undergone adaptive changes during the divergence of the two species.

One way around this problem involves knowing where to look for adaptive evolution. Lysozyme in colobine monkeys was always likely to be a good candidate but, even there, comparisons of the entire gene sequence between extant species reveal adaptive change with low efficiency³. Alternatively, if we knew beforehand that specific sites within a protein were candidates for adaptive changes, we could concentrate on these and exclude many positions subject to purify-

Coding DNA substitutions

Many molecular evolutionary analyses, particularly those attempting to detect adaptive evolution, rely on distinguishing between synonymous and non-synonymous differences in DNA sequences (see ref. 8). A synonymous (or 'silent') substitution is one that, owing to the degeneracy of the genetic code, makes no change to the protein sequence encoded; a non-synonymous substitution results in an amino-acid replacement. The extent of each type of change can be estimated as K_A and K_S , respectively the numbers of non-synonymous and synonymous substitutions per site. For any comparisons other than those between very closely related sequences, these values

have to be subjected to statistical methods to attempt to correct for multiple hits; that is, successive substitutions superimposed at the same site.

The approach that has been widely applied in the search for adaptive evolution of proteins is to contrast these non-synonymous and synonymous substitution rates, using the ratio K_A/K_S . Because synonymous changes seem (at least in many genes, and particularly those of mammals) to be largely neutral, and because population genetics theory indicates that neutral substitutions should occur at the neutral mutation rate, this K_A/K_S ratio is in effect examining whether non-synonymous substitutions

have occurred faster or more slowly than this neutral rate. Most non-synonymous mutations are likely to be deleterious and therefore unlikely to become substitutions by spreading to fixation in the population. Thus, much of the natural selection at non-synonymous sites is expected to be negative or purifying, leading to a reduction in K_A below the neutral rate (and so to K_A/K_S ratios less than one). But if a non-synonymous mutation is advantageous, its process of fixation can happen much faster than the neutral rate. So a K_A/K_S ratio greater than one is taken as indicating that positively selected changes have inflated the numbers of non-synonymous substitutions. P.M.S.

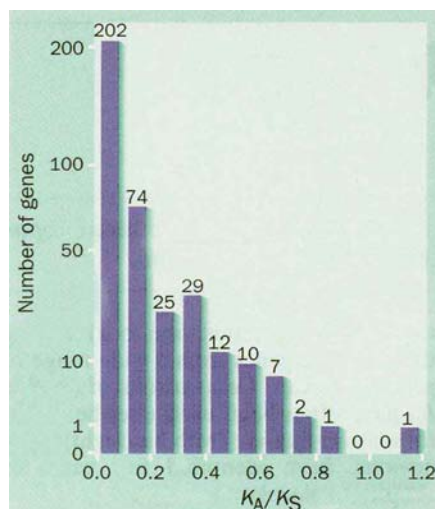


FIG. 2 Ratios of non-synonymous/synonymous divergence (K_A/K_S) for 363 genes compared between mouse and rat⁴. The median value is 0.08. Only one gene (that encoding interleukin-3) has a value greater than one.

ing selection. For example, armed with the knowledge of which residues within class I major histocompatibility proteins form part of the antigen-recognition site, it was possible to show that these sites specifically have an increased K_A/K_S ratio, whereas other codons within these genes do not⁵. Unfortunately, comparable prior information for other proteins is scarce.

Another solution involves being able to invoke adaptive change even when the K_A/K_S ratio is less than one. Under the neutral theory (that is, if all substitutions occurred through random genetic drift without the assistance of natural selection) the ratio should be the same for intra- and interspecific comparisons of sequences. Thus, even if the ratio is less than one, a significantly higher ratio between species than within can indicate that (at least some of) the changes between species reflect mutations that were favoured by selection. This, in essence, was the basis of a test introduced by McDonald and Kreitman to identify adaptive changes in alcohol dehydrogenase among *Drosophila* species⁶.

The McDonald–Kreitman test can only be used for very closely related species. Genes sharing a recent common ancestry have only a few synonymous substitutions, and so if a number of non-synonymous changes have occurred, there is a better chance that they can be flagged as being

adaptive (that is, produce a K_A/K_S ratio greater than one), even in the absence of intraspecific data. Clearly, these tests can only detect adaptation that has happened recently.

In contrast, Messier and Stewart have been able to infer ancient periods of adaptive change using a phylogenetic approach. Phylogenetic analyses attempt to apportion the evolutionary differences between extant sequences to branches within the evolutionary tree that connects them. It is possible to estimate the sequences of common ancestors at nodes within the tree, and perform the K_A/K_S analysis on comparisons between these ancestral sequences. Thus, success has come from the same general principle, comparing closely related sequences among which few synonymous substitutions have occurred, but this time using sequences that have long been extinct.

Colobine monkeys as a whole are leaf-eaters, so not surprisingly most of the adaptive substitutions in lysozyme seem to

have occurred during the origin of the group. It is a surprise, however, that the analysis also points to a burst of adaptive substitution in lysozyme during the early evolution of hominoids, around 20 million years ago. What selective pressures drove these changes, and how they have produced adaptations, remain unknown.

The technique used by Messier and Stewart could be applied to any gene; the requirements are sequences from enough species to produce a densely branching tree. But which genes should we be looking at? Darwinian evolution is best epitomized by numerous examples of morphological adaptation. To find their genetic bases, we may need to pay especial attention to regulatory factors, such as genes controlling development, or regulatory sequences controlling gene expression⁷. □

Paul M. Sharp is in the Department of Genetics, University of Nottingham, Queens Medical Centre, Nottingham NG7 2UH, UK (e-mail: paul@evol.nott.ac.uk).

COSMOLOGY

The Universe as a lattice

Robert Kirshner

WHAT are the largest structures in the Universe? On page 139, Einasto *et al.*¹ report the latest attempt to elucidate the organization of the Universe on the 100-million-parsec scale, by looking at the three-dimensional distribution of galaxy clusters. Startlingly, they see hints of a pattern: although the data are sketchy and the interpretation depends on details of the catalogue of clusters used, the network of superclusters and voids seems to form a three-dimensional lattice with a size of $\sim 120 h^{-1}$ Mpc (where h^{-1} is the Hubble constant in units of $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$).

The largest scale for structure is an important number, as it is determined by the state of the Universe at the time when hydrogen atoms first combined from the ionized soup. The pattern of fluctuations seen today constrains conditions in the Big Bang and illuminates physical events that unfolded when the Universe was a thousandth of its present age. One route to this hidden era of structure formation is studying the pattern of irregularities in the cosmic microwave background². Another is to stretch our local map of galaxy positions to ever-larger scales, as Einasto and colleagues have done.

At the moment, we have only measured the microwave background on huge scales that don't connect with the galaxy observations. Roughly speaking, the well-measured background fluctuations correspond to a scale of 2,000 Mpc, while the galaxy measurements that have much statistical grip are just now reaching beyond 200 Mpc.

The theorists try to bridge that gap with huge numerical simulations. They show how gravity makes the density contrast increase from soup to chowder for any particular original spectrum of fluctuations in the underlying dark matter. If all goes well, the lumps we see as 10^{-5} variations in the microwave background will grow to reproduce the richly textured, high-contrast Universe of galaxies and galaxy clusters we see today.

The investigation by Einasto *et al.* is the latest of a number of clues, none very compelling by itself, that hint at the presence of structure in the Universe on the 100-Mpc scale. For example, redshift surveys show that galaxies are certainly gathered into a pattern of high-density walls and low-density voids^{3–5} on scales of 50 to $150 h^{-1}$ Mpc.

The problem is that observers have a finite amount of telescope time, and so must choose whether to survey many galaxies over a large angle to moderate depth, or to concentrate on a small region and measure the relative distances to faint, distant galaxies. One early and provocative result⁶ using the latter, 'pencil-beam', method showed periodic peaks in the galaxy distribution with a scale of $128 h^{-1}$ Mpc. But because the survey area was so narrow, it was hard to tell whether this was a fundamental feature of the galaxy distribution, or just a somewhat misleading way of visualizing nearby walls and voids.

Since then, larger redshift surveys have pushed out the volume in which structure could be seen in detail, by diligence and

- Haldane, J. B. S. in *On Being the Right Size* (ed. Maynard Smith, J.) 13–25 (Oxford Univ. Press, 1985).
- Stewart, C. B., Schilling, J. W. & Wilson, A. C. *Nature* **330**, 401–404 (1986).
- Messier, W. & Stewart, C. B. *Nature* **385**, 151–154 (1997).
- Wolfe, K. H. & Sharp, P. M. *J. Molec. Evol.* **37**, 441–456 (1993).
- Hughes, A. L. & Nei, M. *Nature* **335**, 167–170 (1988).
- McDonald, J. H. & Kreitman, M. *Nature* **351**, 652–654 (1991).
- Jenkins, D. L., Ortori, C. A. & Brookfield, J. F. Y. *Proc. R. Soc. Lond. B* **261**, 203–207 (1995).
- Li, W.-H. & Graur, D. *Fundamentals of Molecular Evolution* (Sinauer, Sunderland, MA, 1991).

by new techniques that use optical fibres to channel the light from each source to the spectrograph, thereby multiplying the effective observing time by a factor of 100 or more.

The power spectrum is a useful measure of the amount of density variation on each scale (see Fig. 2 on page 140). The power in galaxy fluctuations grows as the scale increases^{7,8} up to about 100 Mpc. Similarly, the power spectrum of fluctuations in the microwave background rises towards shorter wavelengths than the COBE scale. The maximum must lie in the poorly observed range from 200 to 2,000 Mpc.

But Einasto *et al.* hint at something more than just incoherent power. Rather, they see a structure they describe as a rectangular three-dimensional lattice. This is not unprecedented — the analysis of the Las Campanas redshift survey by Landy *et al.*⁹ also gave a hint of periodic clustering on scales of about 100 Mpc — but it is puzzling. What could be the origin of extra power at a particular large scale? One exotic possibility is new physics that would alter the spectrum of perturbations emerging from the conjectured period of inflation. A less daring suggestion, made by Alex Szalay of Johns Hopkins University, among others, is that acoustic waves in the early Universe leave their imprint on the cosmic background radiation.

This 'Doppler peak' in fluctuation is the observational target of two new satellite projects: MAP and COBRAS/SAMBA. For the right combination of cosmic parameters, with fashionably low values of the cosmic density, the peak produced by acoustic waves comes out near 100 Mpc, and it may set the stage for the formation of the structures that are dimly glimpsed in the redshift data.

Fortunately, a new era of precision cosmology is just around the corner. Massive redshift surveys, such as the '2DF' survey at the Anglo-Australian Telescope and the Sloan Digital Sky Survey, will make a great improvement in the measurement of galaxy structures out to 500 Mpc. When combined with forthcoming microwave-background measurements, these probes of cosmic structure will show whether the recent hints of super-large organization are signs of real events in the early Universe. □

Robert Kirshner is at the Center for Astrophysics, Harvard University, 60 Garden Street, Cambridge, Massachusetts 02138, USA.

Betas are brought into the fold

Brent L. Iverson

THE research groups of chemists Samuel Gellman¹ and Dieter Seebach² have independently reported synthetic molecular chains that can fold into a surprisingly stable new type of structure analogous to the α -helix. Instead of the α -amino acids (1 in Fig. 1) found in natural proteins, the Gellman and Seebach systems are composed of linked β -amino-acid units (2).

This has brought into focus recent efforts to design synthetic molecules that fold into well-defined three-dimensional structures. A goal of this emerging field is to extend the repertoire of folded molec-

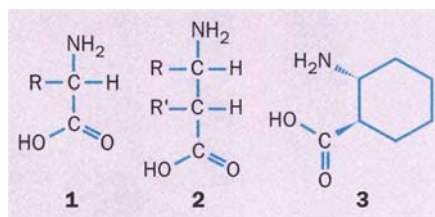


FIG. 1 α -(1) and β -(2,3) amino-acid units.

ular structures beyond those that exist in nature. The new work demonstrates in spectacular fashion that chains of naturally occurring α -amino acids or nucleic acids do not have the monopoly on folded structures, and that molecular engineers are not limited to the folding motifs found in nature^{3,4}.

The larger biological molecules, especially proteins and RNA, owe their complex activities to stably folded structures that result largely from a balancing of non-covalent interactions. Attractive, repulsive, solvation and entropic forces strike a dynamic balance that enables the flexible chains to fold into just a single functional protein or RNA conformation, out of the astronomical number of possibilities. It is this folding that turns the one-dimensional information of the genetic code into the three-dimensional molecules that are responsible for life. So before synthetic molecules can be designed and produced that approach the exquisite complexity of natural proteins or RNA molecules, these same forces need to be harnessed in predictable ways.

Gellman and co-workers used nuclear magnetic resonance, circular dichroism and X-ray crystallography to confirm that chains composed of six β -amino-acid units, in this case *trans*-2-aminocyclohexanecarboxylic acid (3 in Fig. 1) form well-defined helices in methanol solution and in the solid

state¹. As predicted by computer simulations, the helix is stabilized by hydrogen bonds between every third unit. This folding pattern results in 14-atom 'rings' being formed by the hydrogen bonds, so Gellman has called the structure the 14-helix. The 14-helix has a length of 5.0 Å per turn, and an almost perfect three-fold symmetry owing to its having 3.0 residues per turn. The standard α -helix found in proteins has 3.6 residues and 5.6 Å per turn.

The authors used NMR to measure the time it took for amide hydrogen atoms to exchange with hydrogens from the solvent. The amide hydrogens usually exchange within minutes unless they are taking part in strong hydrogen bonds, but the two interior hydrogen-bonded NHs in Gellman's 14-helix took more than *two days* to exchange. These extremely slow exchange rates demonstrate the remarkable stability of this folded structure. In contrast, α -amino-acid peptides in this size range rarely show any identifiable folding in solution, much less a robust structure.

The cyclohexyl rings presumably help to stabilize the structure by restricting the conformational flexibility of the individual β -amino-acid units. The rings radiate out away from Gellman's 14-helix to form three 'stacks' (Fig. 2). In addition, they are ideal sites for adding groups to make derivatives of the basic 14-helix.

Earlier last year, Seebach *et al.*² described β -amino-acid analogues prepared in two synthetic steps from the corresponding α -amino acids. In this case, a $-\text{CH}_2-$ group was inserted between the α -carbon atom and carboxylic acid of each α -amino acid. Peptides resulting from these β -amino acids, called β -peptides, were investigated for folding. One of them showed behaviour consistent with its being highly structured in solution. NMR spectra were used to construct a model of the structure in pyridine, and once again

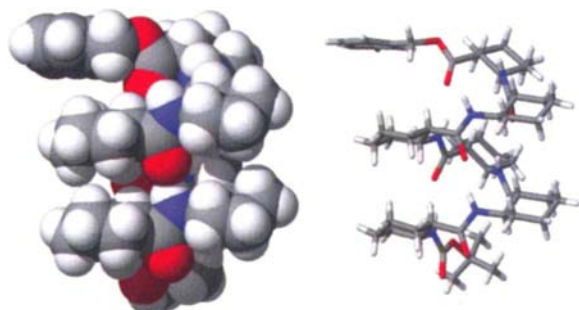


FIG. 2 Space-filling and line representations of Gellman's 14-helix¹, showing the periodicity and consequent stacking of cyclohexyl units.

- Einasto, J. *et al.* *Nature* **385**, 139–141 (1997).
- Smoot, G. F. *et al.* *Astrophys. J.* **396**, L1–L5 (1992).
- Geller, M. J. & Huchra, J. P. *Science* **246**, 897–903 (1989).
- Kirshner, R. P., Oemler, A., Schechter, P. L. & Shectman, S. A. *Astrophys. J.* **248**, L57 (1981).
- da Costa, L. N., Vogeley, M. S., Geller, M. J., Huchra, J. P. & Park, C. *Astrophys. J.* **437**, L1–L4 (1994).
- Broadhurst, T. J., Ellis, R. S., Koo, D. C. & Szalay, A. S. *Nature* **343**, 726–728 (1990).
- Vogele, M. S., Park, C., Geller, M. J. & Huchra, J. P. *Astrophys. J.* **391**, L5–L8 (1992).
- Lin, H. *et al.* *Astrophys. J.* **471**, 617 (1996).
- Landy, S. D. *et al.* *Astrophys. J.* **456**, L1–L4 (1996).

the data were consistent with a 14-helix.

It is worth pointing out that the Seebach system does not have the cyclohexane rings of Gellman to restrict flexibility of the β -amino-acid units. So the presence of identifiable secondary structure in such a small β -peptide is extraordinary, especially when one considers that the extra $-\text{CH}_2$ -group of the β -amino acids might be expected to make the resulting β -peptides more flexible than α -amino-acid peptides, not more structured.

The groups of Gellman and Seebach have established that α -amino acids are not the only amino-acid building blocks that can form molecular chains with interesting folded structures. Their work is a nice complement to the growing number of studies using α -amino-acid building blocks⁵, sometimes in conjunction with abiotic elements^{6,7}, to create novel folded structures. There is also a small but expanding list of entirely abiotic systems that use interactions such as metal binding (the helicates, for example, as reviewed in ref. 8) or stacked aromatic units (such as the aedamers⁹ or molecular ribbons¹⁰) to stabilize folded structures. A critical next step will be to 'bundle' abiotic folded ele-

ments into even larger, more complex structures that contain cavities in which binding and/or catalysis can take place.

The lessons from nature are clear: complex molecular structure can be achieved through the folding of flexible chains. Chemists are now learning how to fold synthetic molecules in a predictable manner. Such efforts can be justified in the short term on aesthetic grounds, but perhaps in the long term on practical grounds as well. Biological chemistry is precise by any measure — molecules are made in living systems essentially free from the unwanted side-reactions that often plague synthetic reactions in the laboratory. Nature achieves this precision through recognition. Substrates must bind in precise ways to nature's catalysts, namely the enzymes, before reacting.

Precise recognition requires a relatively large binding pocket, and that leads back to the issue at hand. Perhaps, one day, the large structures needed to achieve precise, programmable recognition will be produced through the folding of synthetic chains. If this recognition can be combined with efficient catalytic chemistry, then maybe a new generation of molecular machines can be created to help bring the precision level of *in vitro* chemistry closer to that currently only available *in vivo*. Whatever lies ahead down the road for synthetic molecules that fold, the journey promises to be a particularly challenging and interesting one. □

Brent L. Iverson is in the Department of Chemistry and Biochemistry, University of Texas at Austin, Austin, Texas 78712, USA.

1. Apella, D. H., Christianson, L. A., Karle, I. L., Powell, D. R. & Gellman, S. H. *J. Am. Chem. Soc.* **118**, 13071–13072 (1996).
2. Seebach, D. et al. *Helv. Chim. Acta*, **79**, 913–941 (1996).
3. Ball, P. *Designing the Molecular World* (Princeton Univ. Press, 1994).
4. Collman, J. P. et al. in *Organic Synthesis Today and Tomorrow* (eds Trost, B. M. & Hutchinson, C. R.) 29–45 (Pergamon, Oxford, 1981).
5. Bryson, J. W. et al. *Science* **270**, 935–940 (1995).

6. Tuchscherer, G. & Mutter, M. *J. Biotechnology* **41**, 197–210 (1995).
7. Nowick, J. S., Mahrus, S., Smith, E. M. & Ziller, J. W. *J. Am. Chem. Soc.* **118**, 1066–1072 (1996).
8. Amabilino, D. B. & Stoddart, J. F. *Chem. Rev.* **95**, 2725–2828 (1995).
9. Lokey, R. S. & Iverson, B. L. *Nature* **375**, 303–305 (1995).
10. Breidenbach, S., Ohren, S. & Vögtle, F. *Chem. Eur. J.* **2**, 832–837 (1996).

CRYSTAL STRUCTURE

How hard spheres stack up

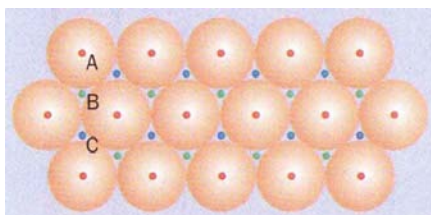
Roberto Car

WHEN stacking atomic spheres like cannonballs, the two simplest arrangements are the face-centred cubic (f.c.c.) and the hexagonal close-packed (h.c.p.) structures. Why do some substances choose to crystallize one way rather than another? Even in a model in which the atoms behave like impenetrable spheres that exert no forces on one another when not touching, the question of which of these two crystalline phases is thermodynamically the more stable has for a long time defied the scientific community. But that may no longer be so: computer simulations by L. V. Woodcock (page 141 of this issue¹) indicate that the f.c.c. phase is favoured over the h.c.p. phase by a tiny, but significant, entropy difference.

Both structures result from the periodic stacking of layers of identical atomic spheres that are close-packed in a triangular lattice, but they have different stacking sequences (see figure). The local environment of each atomic sphere in the h.c.p. and f.c.c. structures is very similar because the first two shells of neighbouring spheres are identical in the two structures. Differences only occur in the third and successive neighbours. The two structures have the same density, with a packing fraction of 74%.

At finite temperature, this close-packing condition is only realized when the pres-

sure is infinite; in general, the kinetic energy of thermal motion causes the spheres to collide like billiard balls, reducing the packing fraction. When it is reduced to about 54%, the hard-sphere solid, either f.c.c. or h.c.p., becomes unstable, and the balls start to diffuse through-



A two-dimensional array of close-packed balls centred at the 'A' sites of a triangular lattice. A second identical layer can be stacked on top of this one with the centres of the balls above either the sites B or C. A similar alternative is possible for each additional stacked layer. The stacking sequence ABABAB is the h.c.p. structure; the sequence ABCABC is f.c.c.

out the entire volume of the sample, signalling a transition to the fluid state. The transition is characterized by a jump of the packing fraction from about 54% in the solid to about 49% in the fluid state. By reducing the pressure still further after the transition has taken place, the fluid gradually and continuously becomes an ideal gas.

This hard-sphere system is probably the simplest model that reproduces qualitatively the characteristics of the melting transition of real systems. Its thermodynamic properties depend uniquely on geometrical (packing) considerations. Indeed, geometrical considerations are the basis of the melting transition in non-associated liquids, and in the case of rare gases like argon the melting properties deduced from the hard-sphere model agree even quantitatively with the experimental data. That is because the attractive part of the inter-atomic potential is relatively unimportant at melting, which is dominated instead by the repulsive part².

In thermodynamics terms, what matters is the Gibbs free energy — the energy available for conversion to work in a reversible transformation at constant pressure and temperature. Work cannot be extracted from a stable phase (one that has the lowest free energy of all competing phases), so to locate a melting transition, the relative free energies of the fluid and solid phases are needed. This is not at all trivial, even for the 'simple' hard-sphere model system. The trouble is that free energies are not directly accessible to computer simulations, which allow us to compute only the free-energy difference between two states that are very close in pressure–temperature space. Computing an absolute free energy means integrating that free-energy difference along a suitable path that connects reversibly the system in its actual state to a reference state of known free energy. That often requires a combination of physical

ingenuity, sophisticated computer simulations and analytical techniques. In this way, the relative stability of the solid and the fluid phases for a hard-sphere system can be determined³.

In principle, the relative stability of the f.c.c. and the h.c.p. phases could be found in a similar way. But it is much more difficult, because the f.c.c. and the h.c.p. phases are so similar and their free-energy difference is so small that exceedingly high numerical accuracy is needed to obtain a definite answer to this question. Woodcock finds that f.c.c. has a lower free energy than h.c.p. by only $0.005RT$ (R is the universal gas constant) within a 20% uncertainty.

This value is smaller than the numerical accuracy of previous calculations, which explains why determining the relative stability of f.c.c. and h.c.p. hard-sphere phases has been an elusive challenge for such a long time. The free-energy difference comes from a small region of the phase diagram around the melting transition, where a few inter-particle collisions occur between spheres more distant than second neighbours. These are different in f.c.c. and h.c.p. structures (see Table 1 on page 143). Because the collisions turn out to be more frequent in h.c.p., at a given pressure it has a slightly lower density than f.c.c., and melts at a higher pressure. This small difference in packing fraction can be ascribed to a small difference in entropy.

What does this result mean for the relative stability of f.c.c. and h.c.p. in real substances? The real systems closest to hard spheres are the rare gases; and, indeed, at low temperature all the rare gases are f.c.c., with the notable exception of helium, which at low temperature and at large external pressure, is h.c.p. This might seem puzzling, because helium is the most hard-sphere-like of all the rare gases. But it is also a very light atom, for which quantum effects at low temperature are very important.

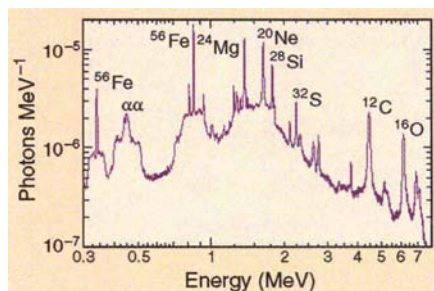
At low temperature, Woodcock's free-energy difference tends to zero. But a simple quantum calculation, in which each atomic sphere behaves like a free particle confined within its primitive lattice cell, indicates that the energy of zero point motion should be larger for f.c.c. than for h.c.p., favouring h.c.p. stability at the lowest temperatures⁴. At higher temperature, the classical result takes over and the system becomes f.c.c.. Such a transition is observed experimentally in helium. □

Roberto Car is in the School of Physics, University of Geneva, 24 Quai Ansermet, 1211 Geneva, Switzerland.

In line for a new mission

Peter J. T. Leonard

THERE is a great wealth of astrophysical information to be gleaned from low-to-medium-energy γ -rays (between about 30 keV and 30 MeV), on a variety of phenomena including solar flares, supernovae and the mysterious cosmic γ -ray bursts. The latter emit most of their flux at these energies. These and other scientific topics were the subject of a meeting late last year*, as were the technologies from which to build telescopes sensitive enough to study them. I will limit myself to just



A theoretical nuclear de-excitation spectrum, showing both broad and narrow emission lines for material of solar composition. (Provided by Reuven Ramaty, NASA Goddard Space Flight Center.)

one topic: nuclear emission lines.

The most direct way to measure the abundances of different elements and isotopes is through their nuclear emission lines. From a solar flare in 1991, for example, OSSE (the Oriented Scintillation Spectrometer Experiment) on board the Compton Gamma-Ray Observatory (CGRO) saw lines from ^{56}Fe , ^{24}Mg , ^{20}Ne , ^{28}Si , ^{12}C , ^{16}O and ^{15}N . These nuclei get excited by energetic protons and α -particles, accelerated to MeV energies by the flare; and they promptly de-excite by emitting γ -rays at characteristic energies. The relative strength of each line can be used to measure isotopic abundances in the Sun without resorting to assumptions about temperature or atomic physics.

The same MeV protons and α -particles should also create radioactive nuclei in the parts of the Sun near a flare, which will emit γ -rays as they decay. These emissions could be used to study mixing in the solar wind and in the Sun's outer convective layer (R. Ramaty, NASA Goddard Space Flight Center), much as studies of the radioactive fallout from a nuclear explosion can be used to study circulation in Earth's atmosphere. But at present there is no instrument sensitive enough and with enough angular resolution to study this phenomenon in and around the Sun.

Supernovae produce detectable nuclear lines too. Types II and Ib supernovae (SNe II and Ib) occur when the cores of massive, evolved stars collapse into neutron stars, and the expanding remnants spread newly synthesized heavy elements into the interstellar medium. Shortly after the launch of CGRO, OSSE detected the ^{57}Co line at 122 keV in the four-year-old remnant of SN1987A. The abundance ratio of $^{57}\text{Co}/^{56}\text{Co}$ measured by OSSE is significantly smaller than that found by optical observations.

COMPTEL (the COMPTon TELEscope), also on CGRO, detected ^{44}Ti in the supernova remnant Cassiopeia A at 1.16 MeV. Cas A is a mystery, as the available modern evidence suggests that the explosion occurred in the second half of the seventeenth century, and it should have been easily seen at the time — but there is no undisputed record of it. Better γ -ray observations would help solve the mystery of what happened here. And γ -ray observations of a SN II or Ib as it happens should be a great help in understanding these explosions.

Type Ia supernovae (SNe Ia) are an order of magnitude more luminous than SNe II and Ib. They are probably explosions of relatively massive white dwarfs, triggered by accretion from binary companion stars. There is only a small dispersion in the peak luminosity of SNe Ia, which makes them useful extragalactic distance indicators.

COMPTEL barely detected two ^{56}Co lines from the type-Ia supernova, 1991T. These are the strongest of several predicted γ -ray lines, and it would be useful to have a much more sensitive instrument in place ready to observe the next bright SNe Ia. The different models for these explosions (detonation, delayed detonation, deflagration, deflagration plus mixing, and so on) make specific predictions for how the γ -ray spectrum evolves with time, and they can be distinguished with sensitive enough observations (P. Höflich, Univ. Texas, Austin).

In star-formation regions within our Galaxy, ^{26}Al is produced by supernovae and other evolved stars, and decays into ^{26}Mg with a half life of 7×10^5 years. COMPTEL has coarsely mapped this 1.81-MeV emission in the Galactic plane, and provided an upper limit of 350 parsecs on the distance to the Vela supernova remnant. But a more sensitive instrument could see groups of million-year-old supernova remnants and determine the Galactic supernova rate, which would constrain the formation rate of the most massive stars (M. Leising, Clemson Univ.). In addition, the rotation curve

1. Woodcock, L. V. *Nature* **385**, 141–143 (1997).
2. Hansen, J. P. & McDonald, I. R. *Theory of Simple Liquids* (Academic, London, 1976).
3. Hoover, W. G. & Ree, F. H. *J. Chem. Phys.* **49**, 3609–3619 (1968).
4. Alder, B. J., Carter, B. P. & Young, D. A. *Phys. Rev.* **183**, 831–833 (1969).

*Low/Medium Energy Gamma-Ray Astrophysics Mission Workshop, Lansdowne, Virginia, USA, 15–17 November 1996.

Compton telescopes

A COMPTON telescope consists of two separate layers of scintillator material. The upper layer is a converter and tracker made from a low-atomic-weight material; the lower is an imager and calorimeter made from a high-atomic-weight material.

An incoming γ -ray Compton-scatters off one of the atomic electrons in the upper layer, changing its energy and direction. An energetic electron results from the process, and its energy is measured in the upper layer. The γ -ray goes on to the lower layer, where, ideally, it is absorbed, kicking out an even more energetic electron. The energy of the initial γ -ray is then the sum of the two electron energies, and their relative energies also constrain the source of the initial γ -ray to a circle on the sky. Measuring the track of the electron in the upper layer would reduce this circle to a point, tremendously improving the telescope's sensitivity and simplifying the overall analysis.

The next generation of Compton telescopes should be a least an order of magnitude better in sensitivity, angular resolution and energy resolution than anything flown in space before.

P. L.

of the Galaxy can be measured by observing wavelength shifts in bright ^{26}Al sources.

One new satellite is already planned: INTEGRAL will be launched in the year 2001 on a two- to five-year mission (J. Matteson, Univ. California, San Diego). It will carry out high-resolution spectroscopy and imaging in the 15-keV to 10-MeV range. The coded-aperture-mask instruments on INTEGRAL are just about as sensitive as such instruments can be at these energies, but on the aforementioned questions many feel that they will only tease us.

At the meeting, there was a consensus that we should build a large Compton telescope (see box) at least ten times more sensitive than COMPTEL or INTEGRAL, and somehow squeeze it into a Medium-class Explorer (MIDEX) mission to be launched in the next five years. (NASA's goal is to launch roughly one MIDEX per year, each costing less than \$100 million.) This is a challenge, but the new technologies presented at the meeting have the potential to meet it. □

Peter J. T. Leonard is in the Compton Gamma-Ray Observatory Science Support Center, Goddard Space Flight Center, Greenbelt, Maryland 20771, USA.

Myxobacterial bounty

John Mann

THE pre-eminence of taxol* as the cytotoxic drug sensation^{1,2} of the 1990s is about to be challenged by a new natural product — epothilone A. The structure of this novel compound from the myxobacterium *Sorangium cellulosum* was only reported³ in July last year, but already two groups have completed total syntheses^{4,5}.

The alacrity of the response to this synthetic target was prompted by the intriguing similarity between the biological activities of taxol and epothilone A — they both stabilize microtubule assemblies and thus inhibit cell division. They apparently share a common receptor on microtubules, and at least *in vitro* have almost identical cytotoxic activities. In addition, the structure of epothilone is much simpler than that of taxol, because it has seven stereogenic centres altogether (taxol has eleven), and only one macrocyclic ring (four separate rings in taxol).

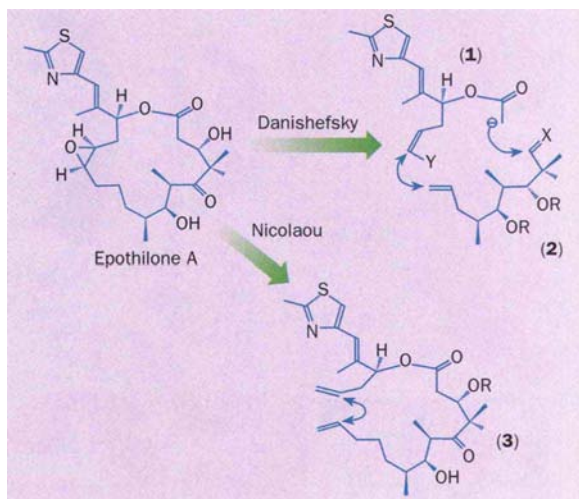
The two new synthetic strategies are entirely different, but both involve ambitious chemistry. Danishefsky and colleagues⁴

prepared the two fragments **1** and **2** (see figure), then coupled the two alkenes via a Suzuki carbon-carbon-bond cross-coupling reaction. They then used the enolate derived from the methyl group of the acetate of fragment **1** to react with a terminal aldehyde of fragment **2**. This macroaldolization, and the subsequent epoxidation to form epothilone A, have a remarkably high stereoselectivity.

Nicolaou and colleagues⁵, in contrast, produced the large intermediate **3** from three smaller fragments, and then produced the key carbon-carbon double bond using a ruthenium-catalysed alkene metathesis process that his group had already developed⁶. Once again, epoxidation completed the sequence. Of the two routes, the latter is marginally shorter and also requires less protection of functional groups. Other research groups are also hot in pursuit of this synthetic target; for example, Schinzer *et al.*⁷ have reported a concise synthesis of key intermediates that should allow access to epothilone.

So is this intense synthetic activity justified? The answer would appear to be 'yes'. Epothilone A is 2,000–5,000 times more active than taxol *in vitro* against certain

cancer cell lines that have multiple drug resistance⁸. But the two real bonuses are its water solubility (about 30 times greater than taxol), and its "almost unlimited availability by fermentation of a technical culture medium"⁹. Although taxol can be obtained from the bark of the Pacific yew, the tree grows extremely slowly so this supply is not viable. A related natural product, 10-deacetyl baccatin III, can be obtained from the needles of the European yew,



Routes to a total synthesis of epothilone A taken by Danishefsky's⁴ and Nicolaou's⁵ groups.

and converted into taxol in four chemical steps. This still means that taxol is an expensive drug. It is likely that epothilone will be produced using fermentation technology, and the chemistry discovered during the total syntheses will be used to prepare semi-synthetic analogues with, one hopes, a better spectrum of activity.

If the compound proves to have good *in vivo* activity, it will become a very important addition to the armoury of cancer chemotherapy. Further, because the myxobacteria have been poorly investigated thus far, the discovery of epothilone offers a tempting glimpse of the treats that might be available from this section of nature's medicine cabinet. □

John Mann is in the Department of Chemistry, University of Reading, Whiteknights, Reading RG6 6AD, UK.

1. Mann, J. *Nature* **367**, 594 (1994).
2. Jenkins, P. *Chem. Br.* **32**, 43–46 (1996).
3. Höfle, G. *et al. Angew. Chem. Int. Edn Engl.* **35**, 1567–1569 (1996).
4. Balog, A. *et al. Angew. Chem. Int. Edn Engl.* **36**, 2801–2803 (1996).
5. Yang, Z., He, Y., Vourloumis, D., Vallberg, H. & Nicolaou, K. C. *Angew. Chem. Int. Edn Engl.* **36**, 166–168 (1997).
6. Nicolaou, K. C. *et al. Angew. Chem. Int. Edn Engl.* **35**, 2399–2401 (1995).
7. Schinzer, D., Limberg, A. & Böhm, O. M. *Chem. Eur. J.* **2**, 1477–1480 (1996).
8. Bollag, D. M. *et al. Cancer Res.* **55**, 2325–2333 (1995).
9. Höfle, G. *et al. J. Antibiot.* (in the press).

*Bristol-Myers Squibb has registered Taxol as a trademark and wishes the scientific community to use the name paclitaxel.

The alphabet of weight control

Jeffrey M. Friedman

In vertebrates — and especially among land-dwelling mammals — the ability to store large quantities of energy-dense fuel in the form of adipose tissue allows survival during prolonged periods of food deprivation. To maintain such fuel stores without undergoing continual alterations in its size and shape, an animal needs to achieve a balance between energy intake and expenditure. A series of reports¹⁻³, including a paper by Fan *et al.*¹ on page 165 of this issue, establish the importance of neuropeptide Y (NPY) and a specific melanocortin receptor as key components of the systems in the brain that regulate body weight. As well as increasing our understanding of the physiological 'wiring diagram' that allows a constant body weight to be maintained, the data indicate that distinct pathways may regulate the biological responses to starvation and to an increase in body weight.

The notion that the brain regulates food intake and body weight dates back to the beginning of this century when it was first observed that, in rare cases, damage to the human hypothalamus results in obesity. The involvement of individual hypothalamic nuclei (collections of neurons) in the regulation of body weight was established in the 1940s by Hetherington and Ranson, who introduced lesions in discrete nuclei within this brain region and found that rodents became obese⁴. These neuroanatomical data, combined with studies that documented the constancy of body weight over time, indicated that energy balance is physiologically controlled by integratory centres in the hypothalamus.

Now the molecular underpinnings of this hypothalamic regulatory system are emerging. Fan *et al.*¹, and Huszar *et al.*² in a study published in this week's *Cell*, have shown that the melanocortin-4 receptor and its peptide ligand, melanocyte-stimulating hormone (MSH), are important in the pathogenesis of obesity in mice that have a colourful mutation known as yellow agouti (*A^y*). And Erickson *et al.*³ have reported in *Science* that mutations in the gene that encodes the brain peptide NPY lead to a reduction in the obesity and other abnormalities that are seen in mice mutated in the *obese* (*ob*) gene.

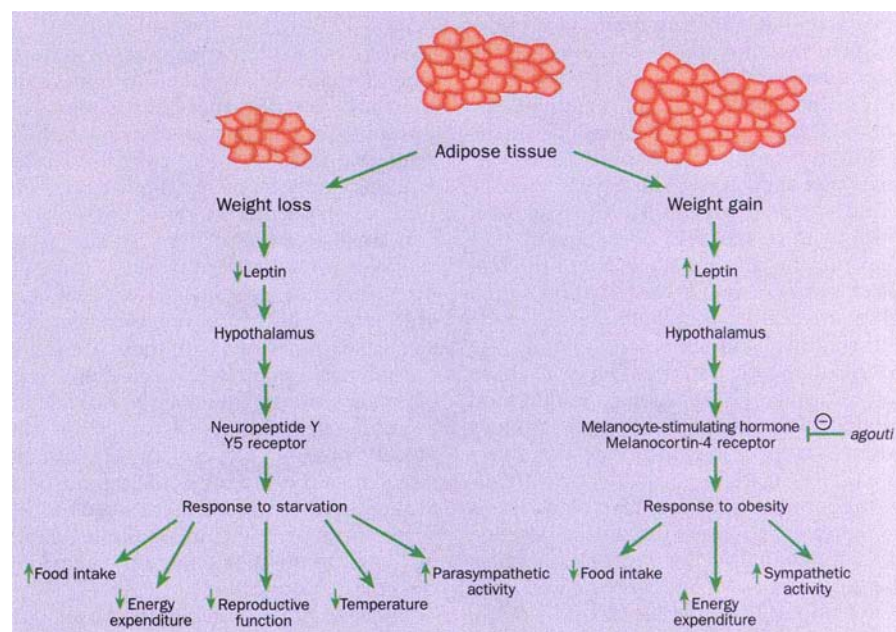
Rather than just confirming that MSH and NPY are involved in the regulation of body fat content, the new data may also help to provide a framework for the system that regulates body weight. This framework can be most simply understood if one considers the physiological responses to weight loss as being functionally distinct from the responses to weight gain (see figure). To develop this

idea further an introduction to leptin is necessary.

Leptin is a hormone that is produced by adipocytes (fat-storage cells). It maintains the stability of body fat mass^{5,6} — a loss of body fat leads to a decrease in leptin, which in turn leads to a state of positive energy-balance. In other words, low leptin levels lead to a state in which food intake exceeds energy expenditure and other adaptive changes are also seen⁷. Conversely, an increase in adiposity leads to an increase in

result from a limited food intake can be prevented⁹.

The study by Erickson *et al.*³ places NPY downstream of a low level of leptin in this biological response. The *NPY* gene was thought to be involved in the development of the *ob* phenotype because injection of NPY into the brain increases food intake and body weight¹⁰. Moreover, the levels of endogenous NPY mRNA are much higher in *ob* than in normal mice, and they decrease following treatment with leptin¹¹. Surprisingly, mice with mutations in the *NPY* gene are of normal weight and they respond normally to exogenous leptin¹². But when Erickson and colleagues bred *NPY* knockout mice



Leptin allows the body to maintain constant stores of fat^{5,6}. A loss of body fat (starvation) leads to a decrease in leptin, which in turn leads to a state of positive energy-balance whereby food intake exceeds energy expenditure, and other responses are also seen⁷. Conversely an increase in adiposity leads to an increase in the levels of leptin and a state of negative energy-balance, whereby energy expenditure exceeds food intake⁷. Fan *et al.*¹ and Huszar *et al.*² have shown that the melanocortin-4 (MC-4) receptor and its ligand, melanocyte-stimulating hormone, are necessary for the biological response to obesity. And Erickson *et al.*³ have bred mice with mutations of neuropeptide Y (NPY) to *ob* mice, and they show that NPY is an important component of the biological response to a low levels of leptin and starvation.

the levels of leptin and a state of negative energy-balance. In this case, energy expenditure exceeds food intake⁷.

In the absence of leptin (as seen in *ob* mice), animals fail to restrain their food intake and they become obese. Such animals have a complex syndrome in which they show most of the signs and symptoms of early starvation — abnormal reproductive function, hypothermia, stunted linear growth, hormonal abnormalities and decreased activity⁸. So *ob* animals seem to exist in a state of perceived starvation and, as a consequence, they become obese. If *ob* mice are treated with leptin, the obesity is corrected and the abnormalities that indicate the starvation response are reversed⁶. Also, if normal mice are treated with leptin, some of the symptoms that

to *ob* mice, they found that NPY is indeed involved in regulating body weight. The *NPY^{-/-} ob/ob* mice were not only less obese than standard *ob* mice, but they also became fertile, were of normal length and they showed an improvement in some of the other abnormalities found in *ob* mice. So NPY is probably a critical component of the biological response to low levels of leptin and starvation. However, the residual obesity seen in the *NPY^{-/-} ob/ob* mice suggests that other downstream factors required for the response to increased weight are still not working normally in these animals. It is intriguing, in this regard, that *NPY^{-/-} ob/ob* mice are phenotypically similar to mice that have an induced mutation in the melanocortin-4 (MC-4) receptor gene.

The sequence of events that led scientists to mutate the MC-4 receptor began with the yellow-agouti mouse mutants. A^Y mice are named for their bright yellow fur but they are also obese^{2,13}. Although A^Y mice are not as obese as *ob* mutants, they still weigh twice as much as controls. Cloning of the *agouti* gene revealed its product to be a secreted factor that is overexpressed in the A^Y mice¹³. When it is overexpressed in the skin, the agouti peptide inhibits the effects of MSH (which is a pigmenting factor) on melanocortin-1 receptors: this leads to the yellow coat colour¹⁴. But this interaction cannot explain why the animals are obese. So it was suggested that the agouti peptide also inhibits the action of MSH on the MC-4 receptor — a transmembrane G-protein-coupled receptor that is expressed in the hypothalamus^{14,15}.

This possibility is now strengthened by the fact that mice with mutations in the MC-4 receptor are as obese as the A^Y mice, yet they do not have yellow coats. Huszar *et al.*² have bred MC-4-knockout mice that are markedly obese, but neither they nor the A^Y mice show any of the features of the response to starvation — the mice are fertile, they have normal body temperature and, if anything, they are longer than normal mice. Fan *et al.*¹ have also shown that MSH agonists inhibit food intake in normal and obese animals whereas MSH antagonists have the opposite effect. In addition, both A^Y and MC-4-knockout mice have very high blood levels of leptin. So it seems that in the absence of a functional MC-4 receptor, animals can no longer respond to increased concentrations of leptin in the blood.

These studies¹⁻³ firmly establish NPY, MSH and MC-4 as components of the alphabet of weight control — they suggest that NPY is an important mediator in the response to starvation, whereas MSH and the MC-4 receptor are key components of the hypothalamic response to obesity (see figure). Although the two limbs of the system may be functionally and anatomically

distinct — NPY and MSH are expressed in different subregions of the brain — there is likely to be extensive crosstalk between them.

Further components of the systems that mediate weight control will probably soon be identified. Many neurotransmitters and neuropeptides are likely to function on each limb of the homeostatic system that regulates body weight. This may have therapeutic implications, and antagonists

of NPY and its hypothalamic Y5 receptor are currently under development¹⁶. So it is fair to expect that the report by Fan *et al.*¹ will not be the last instance in which successful efforts to manipulate food intake pharmacologically are reported. □

Jeffrey M. Friedman is in the HHMI Research Laboratories, Rockefeller University, 1230 York Avenue, New York, New York 10021-6390, USA.

NEUROPSYCHOLOGY

A vision over time and space

Glyn W. Humphreys

'SEEING' the world happens across time as well as space. When asked to introspect, most people report that 'seeing' seems be instantaneous — we have an immediate impression of a full and coherent world. But our impressions are almost certainly false. Highly acute vision is limited to a very small area of the retina, so our representation of a detailed world is built up over time, as we selectively view the objects present. Moreover, even within a single 'snapshot' of vision there are limits on our performance. Enough detail may be extracted from an image to provide an accurate description of some of the objects present, but actions are usually only directed to one object at a time: consequently in the brain there needs to be some form of selection of the available information so that coherent behaviour can emerge over time.

Psychological studies have shown that the process of selection takes a finite and quantifiable time, one measure of which is provided by a phenomenon known as the 'attentional blink'¹⁻³. On page 154 of this issue, Husain and colleagues⁴ report new evidence showing that the duration of this attentional blink is abnormally extended in patients who show symptoms of unilateral visual neglect, which is a relatively common syndrome following stroke, and is associated with damage to the parietal and frontal regions of the right cerebral hemisphere^{5,6}. The work throws new light on how visual processing operates over time as well as space.

Imagine being presented with a rapid stream of letters on a computer screen. All of the letters are in black apart from one, which is in white. You have two tasks to perform — one is to identify the white letter and the other is to decide whether one of the black letters was an X. Each task, when performed independently on the same stimuli, can be easily completed by most people. However, there is a cost involved when the two tasks are combined, and this is what is known as the attentional blink. If the white letter comes first, people find it very difficult to

decide whether there was also a black X if the X occurs within around half a second of the first target letter. This is not a sensory problem — people can detect the black X following the white letter if they do not first have to identify the white letter — it is a central problem reflecting the time taken to select the first stimulus for a response from the other stimuli occurring in the sequence. It is as if the mind cannot attend to a second stimulus whilst it is selecting the first one for a response.

Patients with unilateral visual neglect can fail to respond to stimuli that are presented to the side of space that is opposite (contralateral) to their lesion. Because sensory projections are crossed, a lesion on the right-hand side of the brain affects perception of the visual field to the left, so patients with right-hemisphere lesions and neglect may fail to eat food from the left side of their plate. Neglect is *prima facie* a spatial deficit, because patients fail to respond to stimuli on one side of space. Neglect can also be increased when patients are presented with several stimuli, some on the same (ipsilesional) as well as on the opposite side of space. Under these conditions the ipsi- and contralesional stimuli may be placed in competition for selection into consciousness, and a bias to select ipsilesional stimuli is revealed. So neglect is often thought to provide evidence for selection in vision being fundamentally spatial in nature: selection involves competition between stimuli at different spatial locations^{7,8}.

Now Husain and colleagues⁴ suggest that at least one component of the unilateral neglect syndrome involves the slow selection of visual information over time, rather than over space. They find that patients with unilateral neglect can have an attentional blink that lasts for nearly two seconds and occurs even when all of the stimuli are presented at a single spatial location. Such a prolonged attentional blink is not found in patients with right-hemisphere lesions who do not have

1. Fan, W., Boston, B. A., Kesterson, R. A., Hruby, V. J. & Cone, R. D. *Nature* **385**, 165–168 (1997).
2. Huszar, D. *et al. Cell* **88**, 131–141 (1997).
3. Erickson, J. C., Hollopeter, G. & Palmiter, R. D. *Science* **274**, 1704–1707 (1996).
4. Hetherington, A. W. & Ranson, S. W. *Am. J. Physiol.* **136**, 609–617 (1942).
5. Zhang, Y. *et al. Nature* **372**, 425–432 (1994).
6. Spiegelman, B. M. & Flier, J. S. *Cell* **87**, 377–389 (1996).
7. Maffei, M. *et al. Nature Med.* **1**, 1155–1161 (1995).
8. Coleman, D. L. *Diabetologia* **14**, 141–148 (1978).
9. Ahima, R. S. *et al. Nature* **382**, 250–252 (1996).
10. Zarjevski, N., Cusin, I., Vetter, R., Rohner-Jeanrenaud, F. & Jeanrenaud, B. *Endocrinology* **133**, 1753–1758 (1993).
11. Stephens, T. W. *et al. Nature* **377**, 530–532 (1995).
12. Erickson, J. C., Clegg, K. E. & Palmiter, R. D. *Nature* **381**, 415–418 (1996).
13. Bultman, S. J., Michaud, E. J. & Woychik, R. P. *Cell* **71**, 1195–1204 (1992).
14. Lu, D. *et al. Nature* **371**, 799–802 (1994).
15. Mountjoy, K. G., Mortrud, M. T., Low, M. J., Simerly, R. B. & Cone, R. D. *Mol. Endocrinol.* **8**, 1298–1308 (1994).
16. Gerald, C. *et al. Nature* **382**, 168–171 (1996).

unilateral neglect, so this delay in time over which visual selection operates may be an inherent part of the neglect syndrome. In patients with visual neglect there is an abnormally long period during which the mind is temporarily 'blind' to new stimuli.

The new results indicate that the clinical pattern of neglect can derive from two sources — a spatial bias in selecting visual information (favouring the ipsilesional side) and a prolonged period over which selection occurs. In other words, it takes an abnormally long time for stimuli at the neglected location to be selected by attention, so whilst one process is still being selected, attention cannot be focused on the next stimulus that occurs. It may well be this prolongation of the time for selection that leads to the problems such patients encounter in everyday life, when they fail to respond to objects on the contralesional side. For example, if selection is prolonged for all stimuli and there is also a bias to select an ipsilesional stimulus, then contralesional objects may be missed (particularly if they appear for brief durations). To account for everyday behaviour and associated clinical impairments in that behaviour, it is vital to understand how vision operates over time as well as space, and to acknowledge that temporary mind blindness can be as debilitating as more sensory deficits.

Several questions remain. For example, at present we do not fully understand why visual selection takes time and why this time period may be extended in neurological patients. It could be that time is required to 'bind' visual features together⁷, or it may be that selection is based on synchronized firing of neurons, which requires time to take effect⁹. Brain damage may disrupt feature binding or temporal synchrony in neural firing. The answers should provide us with fundamental knowledge concerning the processes that underlie our seemingly instantaneous experience of the visual world. □

Glyn W. Humphreys is in the Cognitive Science Research Centre, School of Psychology, University of Birmingham, Birmingham B15 2TT, UK.

1. Raymond, J. E., Shapiro, K. & Arnell, K. M. *J. Exp. Psychol. Hum. Percept. Perform.* **18**, 849–860 (1992).
2. Shapiro, K., Raymond, J. E. & Arnell, K. M. *J. Exp. Psychol. Hum. Percept. Perform.* **20**, 357–371 (1994).
3. Duncan, J., Ward, R. & Shapiro, K. *Nature* **369**, 313–315 (1994).
4. Husain, M., Shapiro, K., Martin, J. & Kennard, C. *Nature* **385**, 154–156 (1997).
5. Robertson, I. H. & Marshall, J. C. (eds) *Unilateral Neglect: Clinical and Experimental Studies* (Lawrence Erlbaum, Hillsdale, 1993).
6. Husain, M. & Kennard, C. *J. Neurol.* **243**, 652–657 (1996).
7. Treisman, A. & Gelade, G. *Cogn. Psychol.* **12**, 97–136 (1980).
8. Humphreys, G. W. & Riddoch, M. J. in *Attention and Performance XIV* (eds Meyer, D. E. & Kornblum, S.) 143–162 (Academic, New York, 1993).
9. Singer, W. *Annu. Rev. Physiol.* **55**, 349 (1993).

Microbial breathing lessons

David L. Kirchman

In a recent article in *The New York Times*, Stephen J. Gould barely stifled a yawn over descriptions of evidence of past life on Mars, arguing that we should not be surprised that bacteria could have once flourished on the Red Planet given that they have been so successful on Earth. According to Gould, it has been the Age of Bacteria since life's beginnings about 3,500 million years ago.

Perhaps the report by del Giorgio *et al.*¹, which appears on page 148 of this issue, will surprise even those who are well aware of the preponderance of bacteria on Earth. The authors come to a disarmingly simple conclusion. Respiration by bacteria exceeds plant production in aquatic systems when production is low — that is, carbon going out can sometimes exceed carbon coming into aquatic biota.

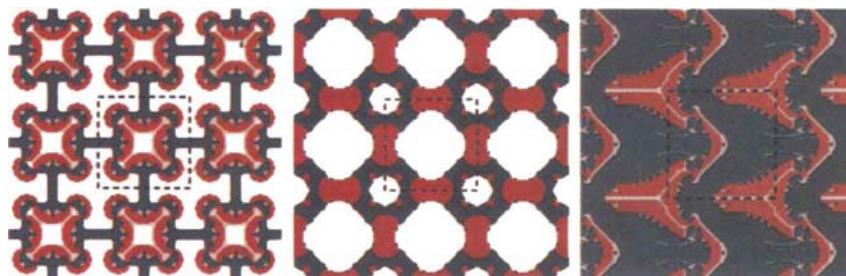
It is hard to imagine an observation with more ramifications for understanding the biogeochemistry of lakes and oceans. For starters, an imbalance between bacterial respiration and plant production would lead to net production of CO₂ and consumption of oxygen, which in turn affects just about everything in nature. Taken at face value, it also implies that there is not enough plant production in

some waters to feed even all bacteria, much less other organisms. If so, it may be easier to understand life on Mars than the existence of life forms higher than bacteria in some of Earth's lakes and oceans.

Explaining how higher life forms can survive is easy for ecologists, once they know what happens to the carbon fixed by terrestrial plants and phytoplankton (primary production). In aquatic environments, much of phytoplankton carbon is fated to be eaten or to sink and to be buried eventually in sediments, but some primary production is also converted to dissolved organic matter (DOM) which is used mostly by heterotrophic bacteria. Uptake of DOM fuels production of bacterial biomass, which in turn supports a microbial food web and eventually larger organisms. Reports of high rates of bacterial respiration² appeared when this world of aquatic heterotrophic bacteria was being discovered some 20 years ago. But microbial ecologists have paid more attention to bacterial production than to respiration, one reason being that it is easier to measure production.

The paper by del Giorgio *et al.* brings our attention back to bacterial respiration by examining published studies of bacteria

Fit-to-shrink materials



CAN a structure made from ordinary materials shrink when heated? Yes, if you add some empty space.

The pattern on the left, made from hypothetical materials of high (red) and low (blue) positive expansivity, has an overall negative thermal expansivity. Each cell scrunches up, with the solids occupying some of the void, and the whole thing gets denser. The volume fractions of each material are fixed.

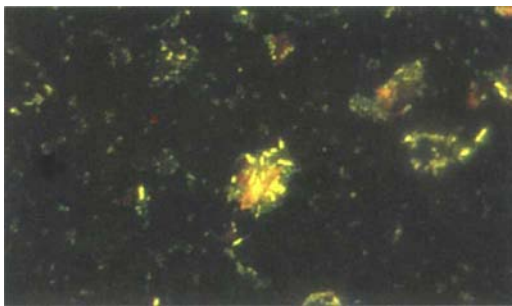
It was designed using an iterative finite-element method (O. Sigmund & S. Torquato *Appl. Phys. Lett.* **69**, 3203–3205; 1996), which can similarly optimize other thermal and mechanical properties. In the centre is a structure designed to expand with as much force as possible. On the right, a stranger shape results from relaxing the geomet-

rical constraints. Here the materials correspond to nickel and Invar. On heating, this pattern contracts with the greatest possible force.

According to the authors' unpublished calculations, arbitrarily high contraction or expansion can be achieved, albeit at the cost of vanishing stiffness. (Such flimsy structures are unlikely to be practical on Earth, but might they find uses in the zero-gravity environment in orbit?)

The designs shown here could be made on the micrometre scale by ordinary lithography and micromachining techniques. And although three-dimensional structures might be trickier to manufacture, the authors plan to design 3D sensors and actuators too.

Stephen Battersby



Small world — a preparation of planktonic lake bacteria, stained with a green nucleic-acid stain and photographed with epifluorescence at $\times 2,360$, showing bacterial colonies and single organisms. (Photograph courtesy of P. del Giorgio.)

and phytoplankton in waters ranging from lakes to the open ocean. They begin by pointing out that there is a significant correlation between bacterial abundance (which is even easier to measure than production) and bacterial respiration in a wide variety of waters; this is an expected observation that lends credibility to their unexpected conclusion. The heart of the paper, though, is discussion of the relationship between bacterial respiration and primary production. It is not surprising that the authors observe a correlation between these two processes. What is surprising is that when primary production falls below roughly $100 \mu\text{g C l}^{-1} \text{d}^{-1}$, bacteria outeat what phytoplankton produce.

Before explaining how any heterotrophic process could exceed primary production, del Giorgio *et al.* defend their conclusion by comparing their bacterial respiration estimates with available data on bacterial production. To do so, they needed to calculate bacterial growth efficiencies, which are a measure of how much organic carbon taken up by bacteria is either transformed into bacterial carbon or respired as CO_2 . The growth efficiencies they calculated range from 10 to 30%, and are lower than that commonly assumed (50%) in ecological models but similar to those directly measured in studies published last year^{3,4}. These figures support the estimates of high rates of bacterial respiration.

The authors go on to hypothesize that growth efficiencies increase with rates of primary production, a view that has many implications for thinking about carbon cycling in lakes and oceans. Although many will argue with del Giorgio *et al.* about the high estimates for respiration, few will find fault with their calculations of low growth efficiencies and the proposal that these efficiencies vary systematically.

The dilemma is then to explain how bacterial respiration could ever be greater than primary production. For lakes, runoff carrying organic material from land (what ecologists call allochthonous carbon) could fuel bacterial respiration independently of phytoplankton production, and indeed del Giorgio *et al.* point out that some lakes

are supersaturated with CO_2 . A similar mechanism is at least conceivable, although debatable, for estuaries and coastal oceans.

But the authors run into real problems with the open ocean, where DOM comes almost entirely from marine sources⁵. They say that highly productive oceanic regions could export DOM to less productive areas and the exported DOM could then fuel high bacterial respiration in waters with low primary production. Although there is some evidence for DOM export⁶, what seems more plausible is that, as del Giorgio *et al.* suggest,

bacterial respiration and primary production are simply not in synch over time.

A specific example may explain this best. During a phytoplankton bloom in the North Atlantic in 1989, primary production was low but bacterial activity remained high on cloudy days⁷. For those days, bacterial respiration exceeded primary production, according to calculations based on the growth efficiencies provided by del Giorgio *et al.* and measured rates of production. The cloudy days, however, were balanced out by sunny ones when primary production was much higher than bacterial respiration. A few days of net heterotrophy may seem more plausible when we remember that phytoplankton release little DOM directly⁸. Grazing and other processes distant in time from phytoplankton are the most likely sources of DOM to support high rates of bacterial respiration when primary production is low.

The new work¹ will be an inviting target for sceptics, as indeed would any study that spans the gamut from nutrient-rich lakes to nutrient-poor open oceans. Even if the imbalance between heterotrophy and autotrophy is eventually righted, bacterial respiration is likely to loom large in future models of carbon cycles. It is becoming clearer and clearer that the Earth remains firmly in the Age of Bacteria — and it will be much easier to revisit the lakes and oceans of del Giorgio *et al.* than to test whether life really did once exist on Mars. □

David L. Kirchman is in the Graduate College of Marine Studies, University of Delaware, Lewes, Delaware 19958, USA.

1. del Giorgio, P. A., Cole, J. J. & Cimleris, A. *Nature* **385**, 148–151 (1997).
2. Williams, P. J. leB. *Oceanol. Acta* **4**, 359–364 (1981).
3. Carlson, C. A. & Ducklow, H. W. *Aquat. Microb. Ecol.* **10**, 9–22 (1996).
4. Cherrier, J., Bauer, J. E. & Druffel, E. R. M. *Mar. Ecol. Prog. Ser.* **139**, 267–279 (1996).
5. Bauer, J. E., Williams, P. M. & Druffel, E. R. M. *Nature* **357**, 667–670 (1992).
6. Peltzer, E. T. & Hayward, N. A. *Deep-Sea Res. II* (in the press).
7. Ducklow, H. W. *et al.* *Deep-Sea Res. II* **40**, 245–263 (1993).
8. Williams, P. J. leB. *Mar. Microb. Food Webs* **4**, 175–206 (1990).

DAEDALUS

God plays dice

THE relativistic and quantum views of the Universe are still irreconcilable. Daedalus now points out the theological implications. If the relativistic view is correct, the Universe is an n -dimensional manifold with time as one dimension. It is a completely defined structure, and everything in it is already determined, future as well as past. God may have declared it good, but there is nothing He or we can do about it now.

The quantum view is quite different. The Universe is steadily emerging from innumerable quantum uncertainties, 'dice thrown by God', as Einstein saw it. Newton himself felt that God must intervene now and then in His creation to keep it on track; and the uncertainty principle provides the ideal mechanism. By tweaking key parameters by less than their Heisenberg uncertainty, He could affect distant outcomes without breaking any physical law. The outcomes, however, might be inconveniently distant. Suppose, for example, that He wished to annihilate the dinosaurs by meteoric collision. He'd have to start the strategy some 100 million years ahead, at which time the required adjustments to the position or momentum of the best available asteroid would be below the Heisenberg limit. But at that time He wouldn't know what sort of dinosaurs would be around — the uncertainty of evolution must multiply up at a much faster rate. If He left the decision till the dinosaurs were already defined, it would be too late to do more than merely shift the aim-point of the asteroid by a kilometre or so.

In other areas, chaos theory allows God to move more rapidly. He could control the weather perhaps a month or two ahead, by tweaking sub-Heisenberg atmospheric fluctuations. And on small highly chaotic systems like the random-noise generators and tumbling balls used to select lottery winners, He could act almost instantly.

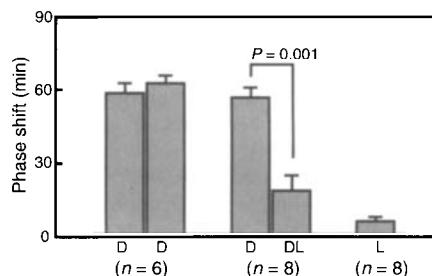
So DREADCO statisticians are studying the records of the British national lottery. It is probably naive to see if the good are significantly over-represented among the winners. God must take a long view, and His social purposes may lie generations ahead. But Daedalus recalls Galton's statistical study of the power of prayer. On average, all lottery players must wish to win equally fervently, no matter how much they have staked. If these wishes affect their chance of winning, a £1,000 punter will not be quite 1,000 times as likely to win as a £1 one. The ratio will be $(1,000+x):(1+x)$, where x is the power of wishing. Daedalus hopes to deduce x from the statistics, and thus discover the influence of human desires on the Divine purpose.

David Jones

Serotonin in the spotlight

SIR — Since the early 1980s, many attempts have been made to use exposure to bright artificial light for the treatment of depression¹. Indeed, for one type of depression, referred to as seasonal affective disorder (SAD) of the winter type, light therapy is now one of the preferred methods of treatment². Because the external light-dark cycle is an important synchronizer of human circadian rhythmicity³, and depression has been associated with disorders of temporal organization^{4,5}, the therapeutic use of artificial bright light for SAD has been suggested to involve the effects of light on biological rhythms. However, the finding that artificial bright light can be effective in treating SAD even when presented in the middle of the day, when light exposure has no effects on circadian phase or day length⁶, is not consistent with the hypothesis that the antidepressive effects of light therapy are mediated by circadian phase changes and/or prolongation of the short days of winter.

The physiological basis for the effective



Treatment with 8-OH-DPAT (D) (1.0 mg kg^{-1} interperitoneally) at circadian time 7 (5 hours before the onset of the subjective night), on two separate occasions, induced comparably large phase shifts in a group of six hamsters whose circadian rhythms were free-running in constant darkness (mean $\pm$ s.e.m.; advances of 58 ± 4 and 62 ± 3 min). In contrast, the phase-shifting effects of 8-OH-DPAT in a similar group of eight animals were significantly attenuated when one of the treatments was combined with a 2-hour light pulse (DL) (56 ± 4 compared with 18 ± 6 min, $P \leq 0.001$). Consistent with the existence of a 'dead zone' for photic phase-resetting of the mammalian circadian pacemaker during the subjective day, exposure to the photic stimulus alone (L) (400–600 lux white fluorescent light) had no effect on circadian phase in an additional group of eight animals (5 ± 2 min).

tive use of light in the treatment of SAD remains unknown. Here we report new data from rodents which raise the possibility that light can significantly modify the processing of serotonergic signals by the mammalian brain.

The selective serotonin re-uptake inhibitors represent one of the most commonly used pharmacological agents for the treatment of depressive illness⁷. In addition to the role of serotonergic mechanisms in the pathogenesis of depression, several animal studies reveal the involvement of serotonin (5-HT) in the control of mammalian circadian rhythmicity. Together with direct and indirect photic input from the retina, the circadian pacemaker in the suprachiasmatic nucleus of the hypothalamus receives serotonergic projections from the midbrain raphe nuclei.

Further, recent studies have demonstrated the ability of several serotonergic stimuli to: (1) advance the phase of the circadian clock when administered during the subjective day⁸, and (2) block the phase-shifting effects of light on circadian rhythmicity during the subjective night (a time when the serotonin stimuli alone have no effect on circadian phase)⁹. The existence of these anatomical and functional connections between retinal and raphe projections to the rodent circadian system provides an excellent model in which to study the effects of light on serotonergic neurotransmission in the mammalian brain.

Here, we used the well-defined ability of the serotonin subtype 5-HT_{1A} agonist, 8-hydroxy-dipropylaminotetralin (8-OH-DPAT), to induce phase advances in the circadian rhythm of locomotor activity in golden hamsters (*Mesocricetus auratus*) during the subjective day. The effects of this stimulus were examined in the same group of animals on two separate occasions, with and without concomitant exposure to a 2-hour light pulse. Consistent with the existence of a 'dead zone' for photic phase-resetting of the circadian pacemaker during most of the subjective day, exposure to the photic stimulus alone had no effect on circadian phase in control animals. However, this intervention considerably attenuated or completely blocked the phase-resetting properties of the 5-HT_{1A} agonist in the experimental group of hamsters (see figure).

The described ability of light to modify the processing of serotonergic stimuli in the circadian system suggests new mechanisms that may underlie the beneficial effects of artificial light therapy in circadian timing and mood disorders. On the

one hand, it indicates that, in the presence of multiple time-giving cues, light exposure can play an important role in circadian phase control even during the 'dead zone' for photic resetting of the circadian pacemaker. On the other hand, it raises the possibility that photic stimulation may, in a similar fashion, alter the functional status of serotonergic pathways outside our circadian model framework — changes which could have important implications for the use of light as a 'drug' to alter neurochemical activity in the brain.

Plamen D. Penev

Phyllis C. Zee

Fred W. Turek

Center for Circadian Biology and Medicine,
Northwestern University,
Evanston,
Illinois 60208, USA

1. Thalen, B. E. *et al.* *Acta Psychiatr. Scand.* **91**, 352–360 (1995).
2. *Diagnostic and Statistical Manual of Mental Disorders* 4th edn 389 (Am. Psychiatric Assoc., Washington DC, 1994).
3. Boivin, D. B., Duffy, J. F., Kronauer, R. E. & Czeisler, C. A. *Nature* **379**, 540–542 (1996).
4. Lewy, A. J., Sack, R. L., Miller, L. S. & Hoban, T. M. *Science* **235**, 352–354 (1986).
5. Van Cauter, E. & Turek, F. W. *Perspect. Biol. Med.* **29**, 510–519 (1986).
6. Wirz-Justice, A. *et al.* *Arch. Gen. Psychiatry* **50**, 929–937 (1993).
7. Andrews, J. M. & Nemeroff, C. B. *Am. J. Med.* **97**, 24S–32S (1994).
8. Lovenberg, T. W. *et al.* *Neuron* **11**, 449–458 (1993).
9. Rea, M. A., Glass, J. D. & Colwell, C. S. *J. Neurosci.* **14**, 3635–3642 (1994).

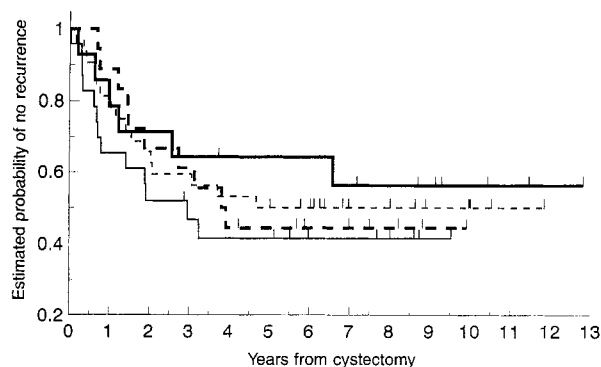
p53 and treatment of bladder cancer

SIR — The report by Waldman *et al.*¹ that tumours with alterations in the gene *p53* may exhibit increased sensitivity to DNA-damaging agents is in significant contrast to the prevailing view that *p53*-altered cells are more resistant to chemotherapy because they are less able to undergo apoptosis (programmed cell death)². This view is supported by clinical studies indicating that *p53*-altered tumours are more resistant to chemotherapy than *p53* wild-type tumours³. The response of *p53*-altered tumours to chemotherapy has profound implications for the selection of patients that would most benefit from such treatment. Here we show that *p53*-altered bladder cancers may show increased sensitivity to chemotherapy that includes DNA-damaging agents.

We have previously shown that patients with transitional cell carcinoma (TCC) of the bladder who demonstrate *p53* alterations have a significantly increased rate of developing tumour metastases and dying of the disease, compared with patients with TCC who have no evidence of *p53* alterations⁴. We have also shown,

Scientific Correspondence

Scientific Correspondence is intended to provide a forum in which readers may raise points of a scientific character. Priority will be given to letters of fewer than 500 words and 10 references.



Kaplan-Meier plots of the estimated probability of non-recurrence for patients with bladder cancer. Solid lines, patients with altered p53; thick lines, patients who received chemotherapy. The difference in recurrence in the p53-altered group is statistically significant ($P < 0.005$); this is not the case for patients with tumours that did not demonstrate p53 alterations ($P = 0.94$).

in a prospective randomized clinical trial, that the administration of adjuvant chemotherapy to patients with TCC resulted in a significant reduction in the rate of tumour recurrences and patient death in the group that received chemotherapy⁵. Because the chemotherapy used in this study included the DNA-damaging agents doxorubicin and cisplatin, we examined these tumours for p53 alterations to determine whether the adjuvant treatment was selectively beneficial in the patients with p53-altered tumours.

Between July 1980 and December 1988, a total of 91 patients with TCC treated by radical cystectomy who had no evidence of systemic metastases were randomized to receive chemotherapy or to be observed (no adjuvant therapy). We examined 88 of these tumours for p53 alterations using immunohistochemical methods, as previously described⁴. The median follow-up period for this group was 8.7 years. Tumours with detectable p53 nuclear immunoreactivity were considered to have p53 alterations; we had previously demonstrated that p53 nuclear immunoreactivity is strongly

The table shows the relative risk of tumour recurrence and death for patients who did not receive adjuvant chemotherapy versus those who received such treatment, stratified by p53 status⁷. The probability of remaining free from recurrence, based on Kaplan-Meier estimates⁸, is plotted in the figure, according to p53 status and treatment received.

These results show that, in patients with tumours that did not demonstrate p53 alterations, adjuvant chemotherapy conferred no recurrence or survival benefit. However, in patients with p53-altered tumours, adjuvant chemotherapy resulted in a threefold-decreased risk of recurrence and a 2.6-fold-increased chance of surviving. Indeed, our results indicate that the only patients to benefit from adjuvant chemotherapy that included DNA-damaging agents were those with p53-altered tumours.

This is the first study in humans to demonstrate that p53 alterations result in increased sensitivity to chemotherapy that includes DNA-damaging agents. The findings are consistent with the mechanisms of chemosensitivity suggested by

Waldman and colleagues¹. Our results are dramatically different from those anticipated by some earlier studies^{2,3}.

Although further study is clearly required, our work suggests that patients with bladder cancer at greatest risk of progression (those with p53-altered tumours) may also benefit most from adjuvant chemotherapy that includes DNA-damaging agents. Fur-

ther, the p53 status of a tumour may help to identify those patients who do not require, and will not benefit from, chemotherapy.

Richard J. Cote*

David Esrig

Susan Groshen

Peter A. Jones

Donald G. Skinner

*University of Southern California School of Medicine,
Kenneth Norris Comprehensive Cancer Center,
1441 Eastlake Avenue,
Los Angeles, California 90033, USA*

1. Waldman, T., Lengauer, C., Kinzler, K. W. & Vogelstein, B. *Nature* **381**, 713-716 (1996).
2. Lowe, S. W., Ruley, H. E., Jacks, T. & Housman, D. E. *Cell* **74**, 957-967 (1993).
3. Sjögren, S. et al. *J. Natl Cancer Inst.* **88**, 173-182 (1996).
4. Esrig, D. et al. *N. Engl. J. Med.* **331**, 1259-1264 (1994).
5. Skinner, D. G. et al. *J. Urol.* **145**, 459-467 (1991).
6. Esrig, D. et al. *Am. J. Pathol.* **143**, 1389-1397 (1993).
7. Pike, M. C. *J. R. Stat. Soc. A* **135**, 201-203 (1972).
8. Kaplan, E. L. & Meier, P. *J. Am. Stat. Assoc.* **53**, 457-481 (1958).

*To whom correspondence should be addressed.

LOWE AND JACKS REPLY — Cote *et al.* raise the interesting possibility that p53 mutation may enhance the chemosensitivity of transitional cell carcinomas of the bladder. This suggestion is in contrast to observations from some experimental systems and from human haematological malignancies, breast carcinoma, and several other tumour types¹⁻⁷, where mutations in the p53 gene have been linked with drug resistance. Clearly, the relationship between p53 mutations and chemosensitivity is of broad clinical significance and may ultimately affect treatment decisions in cancer. Why are the results contradictory?

One difference is methodological. Although the ultimate measure of p53 mutational status is its DNA sequence, Cote *et al.* have instead used immunohistochemical detection of the p53 protein. This technique exploits the observation that p53 protein often accumulates in cells harbouring p53 missense mutations. Why this occurs is unknown, and it is now clear that positive immunohistochemistry is not always indicative of p53 mutations⁸, even in TCC⁹. Wild-type p53 protein also accumulates in response to DNA damage, the expression of certain oncogenes or hypoxia, all circumstances in which elevated protein levels reflect p53 activation rather than mutation. Moreover, not all tumours with missense p53 mutations display positive immunohistochemistry, and those with deletion, frameshift or nonsense mutations would be misclassified as wild type by this method. In clinical studies with limited numbers of specimens, misclassification of even a few tumours can affect the conclusions.

Alternatively, the role of p53 in chemosensitivity could be context-dependent. p53 affects a remarkable number of

RELATIVE RISK OF TUMOUR RECURRENCE AND PATIENT DEATH				
Tumour p53 status	Patients observed	Patients treated*	Relative risk of recurrence [†] (95% CI)	Relative risk of death [†] (95% CI)
No alteration	32	18	1.1 $P = 0.88^{\ddagger}$ (0.8-1.5)	1.0 $P = 0.94^{\ddagger}$ (0.7-1.3)
Altered	24	14	3.0 $P = 0.006^{\ddagger}$ (1.4-6.5)	2.6 $P = 0.005^{\ddagger}$ (1.4-4.7)

Patients with bladder cancer, some of whom were treated with adjuvant chemotherapy, stratified by tumour p53 status. No alteration means no detectable p53 nuclear immunoreactivity; altered means p53 nuclear immunoreactivity was demonstrated.

*Treated with adjuvant chemotherapy. All patients (observed and treated) had undergone radical cystectomy.

[†]Relative risk of recurrence and death of untreated (observed) versus treated patients, after stratifying for lymph-node status, and 95% confidence interval (CI). All deaths, regardless of cause, were counted.

[‡]P values for the comparison of risk of recurrence and death in the observed versus treated groups (stratified log-rank test).

cellular processes, including apoptosis, cell-cycle checkpoints, DNA repair, differentiation and senescence. Perhaps defects in damage-induced checkpoints enhance chemosensitivity, whereas defects in apoptosis promote drug resistance. Consequently, the clinical impact of *p53* mutation could be determined by which effect predominates. This, in turn, may be influenced by tumour type, chemotherapeutic agent, or by other mutations occurring in the tumour cells. The resolution of this issue will require a more complete understanding of the pathways that regulate and are regulated by *p53*.

Although immunohistochemistry is subjective and sometimes unreliable, sequencing is expensive and not without limitations. Even accurate detection of *p53* mutations may not be sufficient for clear-cut clinical correlations. Tumours can retain a wild-type allele or harbour mutations that do not completely inactivate *p53*; thus, tumours identified as 'p53 mutant' may actually display a gradient of activity. Also, different *p53* mutations can have different biological consequences¹⁰, raising the real possibility that a tumour's clinical behaviour varies with mutation type¹¹. Finally, examination of *p53* status alone cannot determine whether the *p53* pathway is intact; tumours with mutations affecting other components of the pathway (for example, *mdm-2*) would be inappropriately classified as wild type.

The results of Cote *et al.*, as well as those linking *p53* mutations to drug resistance, certainly warrant further, larger-scale clinical investigation. Given the issues we have mentioned, however, it could be some time before the complex relationship between *p53* mutations and chemosensitivity is fully understood. It could be argued that the use of *p53* status in clinical oncology does not require a complete understanding of the underlying biology; however, without taking into account the limitations of the current techniques or the complexity of the problem, the broader value of this information is significantly lessened.

Scott W. Lowe

Cold Spring Harbor Laboratory,
Cold Spring Harbor,
New York 11724, USA

Tyler Jacks

Center for Cancer Research and Howard
Hughes Medical Institute,
Massachusetts Institute of Technology,
Cambridge, Massachusetts 02139, USA

1. Lowe, S. W. *et al.* *Science* **266**, 807–810 (1994).
2. Fujiwara, T. *et al.* *Cancer Res.* **54**, 2287–2291 (1994).
3. Bergh, J., Norberg, T., Sjögren, S., Lindgren, A. & Holmberg, L. *Nature Med.* **1**, 1029–1034 (1995).
4. Aas, T. *et al.* *Nature Med.* **2**, 811–814 (1996).
5. Righetti, S. C. *et al.* *Cancer Res.* **56**, 689–693 (1996).
6. Eirouby, S. *et al.* *Blood* **82**, 3452–3459 (1993).
7. Wattel, E. *et al.* *Blood* **84**, 3148–3157 (1994).
8. Sjögren, S. *et al.* *J. Natl Cancer Inst.* **88**, 173–182 (1996).
9. Eslrig, D. *et al.* *Am. J. Pathol.* **143**, 1389–1397 (1993).
10. Rowan, S. *et al.* *EMBO J.* **15**, 827–838 (1996).
11. Taubert, H., Meye, A. & Wurl, P. *Cancer Res.* **56**, 4134–4136 (1996).

Y-chromosome variation in great apes

SIR — A complete lack of variation in the sex-determination gene *ZFY* within species of hominids (great apes and humans) contrasts sharply with the high levels of polymorphism found in mitochondrial DNA sequences from these species^{1,2}. We have sequenced the third intron of the *ZFY* gene in male individuals from the three subspecies of gorilla, the two orang-utan subspecies, and the chimpanzee, and compared these with published sequences from apes and humans. Although there is divergence in the nucleotide sequence between these species, we found no intraspecific variation. Our new evidence suggests that in gorillas, and probably in other great apes, this disparity is due to selection acting on the *ZFY* gene or a closely linked locus.

The paternally transmitted Y chromosome provides a complement to maternally inherited mitochondrial DNA (mtDNA) in phylogenetic analysis, and has been found to mutate at higher rates than autosomes³. In contrast, often surprisingly little intraspecific variation is found at Y-linked loci in humans and other mammals. This discrepancy is usually attributed to either a relatively recent common ancestor⁴, a small effective number of males in the population, or selective sweeps that fix the sequences of linked genes on the nonrecombining portion of the Y chromosome⁵.

A study of polymorphism in the third intron of the *ZFY* gene in a diverse sample of humans uncovered no variation in the 729-base-pair nucleotide sequence⁶. Coalescence theory has been used to estimate the time to a most recent common ancestor as 270,000 years^{6,7}, which is similar to the 200,000 years estimated for the most recent common ancestor of mitochondrial lineages⁸.

We examined the *ZFY* locus in four

individuals of *Gorilla gorilla gorilla* (western lowland subspecies), one *G. g. graueri* (eastern lowland subspecies), and one *G. g. beringei* (mountain subspecies). All were identical, showing no variation within the western lowland populations nor between any of the gorilla subspecies. The same region in two individuals from each of the two orang-utan subspecies, *Pongo pygmaeus abelii* (Sumatra), and *P. p. pygmaeus* (Borneo), also showed no intraspecific variation.

We compared one chimpanzee, *Pan troglodytes*, to the sequence reported for the pygmy chimpanzee, *P. paniscus*⁶. We found only one nucleotide difference (a G→A substitution at position 193) in the *ZFY* locus between these species. The table summarizes recent studies of the rapidly evolving mtDNA control region among gorillas, chimpanzees and humans, together with the *ZFY* data. The other extant hominids display greater divergence between subspecies in mtDNA sequences than the divergence seen among distantly related humans, but just as in humans, no intraspecific variation in *ZFY* sequences has been detected.

The reduced level of polymorphism on the Y chromosome is consistent with selective pressure on *ZFY* or linked loci. Results obtained from applying the HKA test (described in ref. 9) for correlation between polymorphism and divergence at the *ZFY* and mtDNA control region loci are shown in the table. The neutral evolution hypothesis is rejected when two gorilla subspecies are compared, whereas results for gorilla–human comparisons are highly dependent on the effective population size for the *ZFY* locus (which, lacking intraspecific variation, cannot be determined from the data). Alternative explanations for the observed reduction in polymorphism include extensive male

EVALUATION OF THE NEUTRAL EVOLUTION HYPOTHESIS

Comparison	Pairwise nucleotide differences (%)				$N_e^{\text{mtDNA}}/N_e^{\text{ZFY}}$	χ^2	P
	Within population		Between populations				
	ZFY	mtDNA	ZFY	mtDNA			
Gorilla (western)	0	6.9			1	6.7	0.009
versus			0	13.8	3	6.5	0.011
Gorilla (eastern)	–	0.9			10	6.0	0.014
Gorilla (western)	0	6.9			1	4.8	0.03
versus			1.4	18.8	3	2.9	0.09
Human	0	2.5 ⁸					
Gorilla (western)	0	6.9			1	2.7	0.1
versus			1.5	23.8			
Chimpanzee	0.14	11.2 ¹⁰					

HKA test statistic, χ^2 , for evaluation of the neutral evolution hypothesis at the *ZFY* intron and the mtDNA control region^{2,10}. HKA tests using average pairwise differences as previously described⁹. Mutation rates $\mu_{\text{ZFY}} = 0.98$ per sequence per million years (ref. 6), and $\mu_{\text{mtDNA}} = 4.8$ per sequence per million years estimated from the sequence divergence between human and gorilla at the control region locus. Divergence times estimated as 2.5 million years for the gorilla subspecies¹, and 5 million years for the chimpanzee, human and gorilla split⁶. Effective population numbers for the mtDNA locus, N_e^{mtDNA} , were determined by the intraspecific variation at the mtDNA locus, and a generation time of 20 years was assumed.

migration or recent male population bottlenecks. Our limited data set does not allow a robust test of the neutral hypothesis for variation between gorillas and chimpanzees.

Our data indicate that there are factors within the family *Hominidae* that limit levels of intraspecific polymorphism in this region of the Y chromosome relative to the rates of nucleotide substitution noted be-

tween these species. This renders problematic the use of these data in establishing Y-chromosome divergence times.

Wes Burrows, Oliver A. Ryder
Center for Reproduction of Endangered Species,
Zoological Society of San Diego,
PO Box 551, San Diego,
California 92112-0551, USA
e-mail: oryder@jeeves.ucsd.edu

1. Uchida, A. *Trends Ecol. Evol.* **11**, 163–168 (1996).
2. Garner, K. J. & Ryder, O. A. *Mol. Phylog. Evol.* **6**, 39–48 (1996).
3. Tucker, P. K. & Lundrigan, B. L. *Phil. Trans. R. Soc. Lond. B* **350**, 221–227 (1995).
4. Hammer, M. F. *Nature* **378**, 376–378 (1995).
5. Whitfield, L. S., Sulston, J. E. & Goodfellow, P. N. *Nature*, **378**, 379–380 (1995).
6. Dorit, R. L., Akashi, H. & Gilbert, W. *Science* **268**, 1183–1185 (1995).

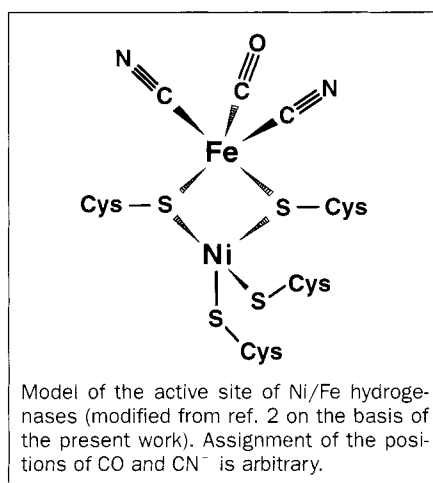
7. Fu, Y. & Li, W. *Science* **272**, 1356–1357 (1996).
8. Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. *Science* **253**, 1503–1507 (1991).
9. Hudson, R. R., Kreitman, M. & Aguade, M. *Genetics* **116**, 153–159 (1987).
10. Kocher, T. D. & Wilson, A. C. in *Evolution of Life: Fossils, Molecules and Culture* (eds Osawa, S. & Honjo, T.) 391–413 (Springer, Tokyo, 1991).

Biological activation of hydrogen

SIR—Hydrogenases are enzymes that can reversibly split molecular hydrogen. Most hydrogen-activating enzymes contain both nickel and iron (termed Ni/Fe hydrogenases), whereas others contain only iron (Fe hydrogenases). In Ni/Fe hydrogenases, the nickel atom¹ and a nearby iron atom² form the site of hydrogen activation. Here we report that one carbon monoxide and two cyanide molecules are also part of this active site. The resulting NiFeCO(CN)₂ prosthetic group is unprecedented in biological systems. The results may be useful for the design of cheap biomimetic hydrogen catalysts for water electrolysis and fuel cells.

Recently, we discovered^{3–5} that Ni/Fe hydrogenases absorb intensely in the 2,150–1,850 cm⁻¹ infrared spectral region, producing a distinct set of bands where no other proteins show absorption. The sensitivity of the absorption bands to changes in the active site indicated to us that the bands might arise from groups (metal-bound diatomic molecules with a triple bond, or triatomic molecules containing two adjacent double bonds) located at the active site of the enzyme⁴. At the same time as these studies, the crystal structure at 2.85 Å resolution² of a Ni/Fe hydrogenase (from the bacterium *Desulfovibrio gigas*) was reported. This structure has an iron atom located within 3 Å of the nickel atom.

Evidence for a low-spin iron ion, distinct from those with high-spin in the iron-sulphur clusters of the enzyme, had



come earlier from Mössbauer studies on the *Chromatium vinosum* enzyme⁶. Nickel and iron are connected by two bridging thiolate groups from cysteine residues. The nickel atom is further anchored to the protein by two more Cys-thiolate ligands. In addition, the electron-density map showed three distinct regions of electron density near the iron atom that were not attributable to protein².

We have now identified the groups responsible for the three anomalous infrared bands using purified *C. vinosum* hydrogenase labelled with either ¹³C or ¹⁵N, by growth of the bacteria on appropriately enriched media. The greater mass of ¹⁵N (compared with natural ¹⁴N) led to a shift to lower frequencies of two of the three infrared bands (see table), consistent with the stretching vibration of diatomic molecules containing one nitrogen atom. The greater mass of ¹³C (compared with natural ¹²C) led to a shift to lower frequencies of all three bands (table). The data indicate the presence of two CN⁻ groups and one CO bound to a metal. Suitable denaturation of the enzyme led to the release of about 0.5 mol CO and 1.6 mol cyanide per

mol enzyme, as determined by published procedures^{7,8}.

We conclude that the three non-protein ligands of iron in the bimetallic active site of Ni/Fe hydrogenases are one carbon monoxide and two cyanide groups (see figure). The resulting strong ligand field around the iron causes this ion to be low spin.

We further conclude that the lone low-spin iron ion observed in Mössbauer spectra of fully-reduced *C. vinosum* hydrogenase⁶ is from the iron atom in the active site. None of the groups detected in the unlabelled enzyme was exchangeable with added ¹³CN⁻ or ¹³CO. The findings are in line with recent observations that the electron density of the non-protein ligands to iron hint at diatomic molecules⁹ deeply buried in the active-site cavity.

Ni/Fe hydrogenases are among the oldest enzymes (3.8 billion years old), demonstrating that early life forms had developed an effective way to activate molecular hydrogen at ambient temperature and pH. Interestingly, infrared bands similar to those in Ni/Fe hydrogenases have also been discovered in two Fe hydrogenases⁵, suggesting a common active-site architecture for all metal-containing hydrogenases.

The involvement of a Ni/Fe binuclear active site containing CN⁻ and CO ligands comes as a surprise. Suitable model compounds incorporating these features have not yet been synthesized or tested. We anticipate that our results will lead to new avenues of enquiry into the development of efficient, cheap and stable hydrogen catalysts for the photovoltaic production of hydrogen by electrolysis of water, or for development of fuel cells, as alternatives to fossil fuels and nuclear energy.

Randolph P. Happe
Winfried Roseboom
Antonio J. Pierik
Simon P. J. Albracht
*E.C. Slater Institute, Biochemistry/FS,
University of Amsterdam,
Plantage Muijdergracht 12,
NL-1018 TV Amsterdam, The Netherlands
e-mail: a311siem@chem.uva.nl*
Kimberly A. Bagley
*Department of Chemistry,
State University College of New York at
Buffalo,
Buffalo, New York 14222, USA*

Infrared absorption bands for hydrogenase			
Enzyme	Band 1	Band 2	Band 3
Unlabelled	2,093	2,083	1,945
¹⁵ N-enriched	2,062	2,050	1,944
Frequency shift	-31	-33	-1
¹³ C-enriched	2,047	2,036	1,900
Frequency shift	-46	-47	-45

Bands are in the 2,150–1,850 cm⁻¹ spectral region (infrared); hydrogenase obtained from *C. vinosum*.

1. Albracht, S. P. J. *Biochim. Biophys. Acta* **1188**, 167–204 (1994).
2. Volbeda, A. et al. *Nature* **373**, 580–587 (1995).
3. Bagley, K. A. et al. *Biochemistry* **33**, 9229–9236 (1994).
4. Bagley, K. A., Duin, E. C., Roseboom, W., Albracht, S. P. J. & Woodruff, W. H. *Biochemistry* **34**, 5527–5535 (1995).
5. Van der Spek, T. M. et al. *Eur. J. Biochem.* **237**, 629–634 (1996).
6. Surerius, K. K. et al. *Biochemistry* **33**, 4980–4993 (1994).
7. Bonam, D., Murrell, S. A. & Ludden, P. W. *J. Bacteriol.* **159**, 693–699 (1984).
8. Nagashima, S. *Anal. Chem. Acta* **99**, 197–201 (1978).
9. Fontecilla-Camps, J. C. *J. Biol. Inorg. Chem.* **1**, 91–98 (1996).

Evolution's evolution

Kevin Padian

Life's Splendid Drama: Evolutionary Biology and the Reconstruction of Life's Ancestry. By Peter J. Bowler. *University of Chicago Press*: 1996. Pp. 525. \$37.95, £30.25.

IN 1887, the palaeontologist H. G. Seeley divided Richard Owen's Dinosauria, named in 1842, into two disparate groups, Ornithischia and Saurischia, based on the structure of the pelvis, vertebrae, braincase and armour. Maintaining that "the Dinosauria has no existence as a natural group", Seeley said that birds, crocodiles, anomodonts (mammal relatives) and pterosaurs would have to be included in the same group, because the characters that united dinosaurs merely "show their descent from a common ancestor rather than their close affinity".

That last curious statement has two implications and, although Peter J. Bowler does not treat this specific case in his excellent new book, it exemplifies much of the post-Darwinian, pre-'modern synthesis' thinking that is his subject. Yes, there are differences between these two subgroups, which were soon accepted as valid. But other similarities noted by Owen indicate a closer relationship between them than to any other animals. 'Different' need not mean 'unrelated'; but what, in the mid-nineteenth century, did 'related' mean, and how could a group have common ancestry but not 'close affinity'? Moreover, if Ornithischia and Saurischia are not most closely related to each other, what other group is?

The answer, as Bowler astutely shows, lies in the transition to true evolutionary thinking — accepting the tree of life in its full ramifications — that was a central problem for disciplines from taxonomy to embryology after the publication of the *Origin of Species*. It was difficult in early post-Darwinian days for scientists to craft a unified approach to what we would now call historical biology. In pre-Darwinian thinking, comparative anatomy and embryology vied for supremacy in explaining the patterns of life, but neither could win: one could not establish homology without the other. Darwin gave homology a new meaning by tying it to common descent, but he learned from Huxley and Owen to mistrust the fossil record as an arbiter of evolutionary history. Palaeontology's ascendancy would come later, with the great discoveries from the American West and elsewhere during the heady decades surrounding the turn of the century. Meanwhile, comparative anatomists and embryologists such as Bateson and T. H. Morgan turned in disappointment to the new science of genetics, and palaeon-

tological theory, in the hands of the Osborns and Schindewolf, languished in the mire of orthogenesis and increasingly implausible land bridges to explain evolutionary and biogeographical trends.

In the study of the history of biology, Bowler's book is greatly needed. It moves the post-*Origin* emphasis away from microevolution (population genetics and so on) and on to macroevolution (the history of life and its environment), which dominated thought in most disciplines until well into the 1900s. Bowler shows that the Darwinian revolution was gradual, not punctuated, and that some disciplines raised its flag well before others did. His first chapter alone, which deals with what Darwin's great book really did and did not do, will set many evolutionary biologists by their ears, and we would all

benefit from taking its perspective to heart. Bowler explored this theme masterfully in some of his earlier books. Here he chooses extended examples of questions from the history of life and examines the forces that shaped the terms of their debate, from pre-Darwinian days to the dawn of the modern synthesis of genetics.

The heart of the book involves that central issue of evolution, the origin and relationships of major groups. Weaving together the three disciplines of comparative anatomy, embryology and palaeontology, Bowler shows how new information became integrated with new ways of thinking about these problems, while the hidebound disciplinary traditions remained. Are the arthropods monophyletic or did they arise from separate stocks? Where did the vertebrates come from? How did they make the transition to land? And what changes were involved in the origins of birds and mammals? Even the reader experienced in one or more of these questions will find new references, forgotten points of view and insights, and earlier expressions of the same arguments advanced today.

This is not, however, a typical effort in



THE buprestids, commonly known as metallic wood-boring or jewel beetles, are popular with collectors because of their metallic streamlined bodies. This plate is taken from *An Inordinate Fondness for Beetles* by Arthur V. Evans and Charles L. Bellamy, an authoritative reference volume resplendently illustrated with line drawings and colour photographs. Holt, \$40.

the 'modern' history of science. While referring to other treatments by historians of science, Bowler's subject is the science itself — specifically, the ideas and disciplines seen against the shifting landscape of explanatory power. But this vantage point can be precarious. History of science has achieved two great advances in recent decades: to abandon the Whiggish view of judging history with the hindsight of present views (no white hats versus black hats), and to approach scientific efforts in the cultural, philosophical and political contexts of their times. Although Bowler almost playfully flirts with some current scientific insight into these questions, for the most part he is thoroughly engaged with the first advance. By contrast, except for a few digressions (such as a tantalizingly elegant analysis of the cultural context of W. D. Matthew's evolving views on biogeography), he avoids the second. To have included it would have made a far different (and far longer) book.

On the other hand, if the cultural context of historically important scientific issues is not explored, and if current views of these issues are not relevant (and Bowler shows effectively that the terms of these questions were entirely different in the past century), what is the contemporary reader to take away? Here history is not so much an object lesson as a snapshot of a parallel world. The failure of scientists to reach consensus on the monophyly of the arthropods, for example, was rooted partly in the patterns of embryological differentiation. The assumption prevailed that embryology was not subject to selection and thus to morphogenetic change. Therefore, because different 'arthropod' groups did not pass through all the same ontogenetic stages, they

could not have come from common origins, and so their adult forms must be convergent. The same kinds of fallacious arguments are raised today, whether the taxon is arthropods or birds, but the fields in conflict are as likely to be systematics and molecular genetics.

For Bowler, as an historian, the patterns are primarily historical, not scientific. The problem is that these questions usually do not attain historiographic resolutions but empirical ones, situated in their contemporary contexts of explanation. Bowler's time frame ends around the 1930s, but not all the questions he explores enjoyed a successful integration of anatomical, embryological and palaeontological traditions by then.

The chapter on arthropods succeeds wonderfully, but those on the origin of vertebrates and the emergence of tetrapods are less successful, and the narratives on bird and mammal beginnings end inconclusively. This is not the fault of Bowler's approach; it is just how history works.

Reading this book, one feels in the hands of a real master of the craft, buoyed by its supple and muscular prose, always confident of the vast expertise behind it. Cavils are minor: some episodes explored here have been studied historiographically not only by historians of science but also by palaeontologists, although the works of the latter are often overlooked. The few typographical errors tend to be concentrated among scientific names that could confuse nonspecialists. On balance, however, this book is one of Bowler's best. □

Kevin Padian is at the Museum of Paleontology, University of California, Berkeley, California 94720-4780, USA.

Bridge to the past

Paul G. Bahn

American Beginnings: The Prehistory and Palaeoecology of Beringia. Edited by Frederick Hadleigh West. *University of Chicago Press: 1996. Pp. 576. \$75, £59.95.*

THE Bering Strait between Siberia and Alaska was named after the Russian explorer Vitus Bering whose expeditions took him there in the mid-eighteenth century. The term Beringia was coined by the botanist Erik Hultén in 1937 to describe only the land bridge that linked the two continents until it was inundated at the end of the Ice Age. But in the 1960s the term was expanded to include north-east Siberia and western Alaska.

This weighty tome certainly lives up to its subtitle, but not necessarily to its title. It comprises an almost encyclopaedic survey, by US and Russian specialists, of the prehistory of Beringia. The editor and the organizers of the different sections must be congratulated for ensuring that the 56 authors produced a series of admirably short articles packed with information. As Yuri Mochanov and Svetlana Fedoseeva put it, "the limited number of pages has forced us to provide descriptions for only the most important and representative relics of the late Palaeolithic".

Roughly a quarter of the book is devoted to palaeoenvironmental studies. It used to be thought that the land bridge was flooded by 14,500 years ago. Newer work has begun to suggest a more recent date of up to 11,000 years ago, although data based on insect remains indicate that a climatic amelioration was already under way by 12,500 years ago. Everybody agrees that the land bridge was a treeless plain. But there is a vigorous and polarized debate between those who, relying on faunal remains, see its landscape as a rich mammoth steppe, and those who, relying on localized pollen samples, believe it was much less productive tundra.

The palaeoenvironmental section is thorough and balanced but it seems bizarre that Dale Guthrie, the leading specialist on Alaska's ancient fauna, is represented only by an abridged article almost 30 years old. And the Russian palynologist Boris Yurtsev, a champion of the steppe tundra concept, is absent, as is Andrei Sher, one of the foremost palaeoecologists of Beringia.

Turning to the far longer archaeological section, it is certainly useful to have short summaries of all the main Alaskan sites gathered in one place, but infinitely more welcome to have similar summaries of the Russian sites available, together and in English. The sites are primarily open-air settlements, with a few caves on either



NATIVE Americans relied on the buffalo as a source of food and spiritual power. But the animal was nearly exterminated in the nineteenth century. In *Buffalo Nation: History and Legend of the North American Bison*, Valerius Geist describes the war waged against the animal by the US army, and its rescue from the brink of extinction in one of the earliest conservation success stories. *Voyageur*, \$35.

side of the divide. Unless one is a fanatic about small stone tools or stratigraphies, these pages are a dull read, but they do form an invaluable work of reference for the prehistoric record of this region.

The book's only major disappointment comes with the two heavily biased concluding articles on "Beringia and New World origins". It is a pity that the editor was not content simply to present a regional record, but felt it necessary briefly to address the vast and thorny subject of the 'first Americans'.

His starting point is that "Clovis is taken to be the basal, the founding, population for the Americas". This is an astounding statement to read today, even allowing for the fact that the book was put together before the results of work in Brazil by Anna Roosevelt and her colleagues were published last April. It is clear from the numerous solid and widely accepted early radiocarbon dates of human occupation sites in South America that it is impossible for the North American Clovis hunters of 11,000 or 12,000 years ago to have been the first people in the New World.

Joseph Greenberg covers the linguistic evidence but promotes his own theory and attacks the alternative views of others, who are not given a voice in this book. The editor's article on the archaeological evidence is even more extreme. He presents the 'pre-Clovis' alternative as resting on three main sites, and sweeps them aside in a few lines. His dismissal of Monte Verde not only rejects the widely accepted date of 13,000 years before present (BP) but also ignores the indisputable human artefacts and features associated with its possible date of 33,000 years BP. His dismissal of Pedra Furada is ill-informed, while his rejection of all the early dates at Meadowcroft as being "quite dubious" is unworthy, as that problem has been dealt with thoroughly and to the satisfaction of all but the most diehard 'Clovis first' adherents.

The editor's attitude even permeates his treatment of the Beringian material. He dismisses the probably very early Russian site of Diring in a couple of lines, claiming it does not "have any bearing on the subject of this volume". And the Yukon's Bluefish Caves — presented quite soberly in the archaeological section, together with their dates of 12,000 years BP and possibly 25,000 years BP — are barely mentioned in his conclusion.

Readers wanting a reference book on the palaeoecology and prehistoric archaeology of Beringia will find this an invaluable resource. But those seeking a balanced and informed view of current debates about the earliest peopling of the New World should look elsewhere. □

Paul G. Bahn writes on archaeology from 428 Anlaby Road, Hull HU3 6QP, UK.

Hazards ahead in the disasters game

David Sumner

Accident and Design: Contemporary Debates in Risk Management. Edited by Christopher Hood and David K. C. Jones. UCL Press: 1996. US distributor Taylor and Francis. Pp. 267. £45, \$75 (hbk); £14.95, \$24.95 (pbk).

Exploring Risk Communication. By Jan M. Gutteling and Oene Wiegman. Kluwer: 1996. Pp. 233. \$118, £79.

Dealing with Risk: Why the Public and the Experts Disagree on Environmental Issues. By Howard Margolis. University of Chicago Press: 1996. Pp. 236. \$27.95, £21.95.

Against the Gods: The Remarkable Story of Risk. By Peter L. Bernstein. Wiley: 1996. Pp. 394. \$27.95, £17.99.

AFTER last November's fire in the Channel Tunnel, a British advertising agency suggested juxtaposing pictures of the incident with those of the Zeebrugge ferry disaster and the Lockerbie air crash, with the caption "Which is the safest way to travel?" But our assessment of risk is hardly ever that simple: it is not just a question of listing deaths.

Our perceptions of risk are inseparable from concepts such as trust and acceptability. And for infrequent but serious events we face the major difficulty of predicting the future from limited past data, a particularly intractable problem where fickle human beings are involved.

Risk: Analysis, Perception and Management, a report produced in 1992 by a Royal Society study group, is a useful resource for those wishing to get up to speed on this subject. But there is tension in this report between the engineers and scientists, who emphasize quantification and tend to view risk as a one-dimensional objective concept, and the social scientists, who stress multidimensionality and the importance of social context. Christopher Hood and David Jones, both of whom were involved in producing the report, extend the debate in their book *Accident and Design*.

The 18 essays in this volume are of variable quality. Some are useful — notably Timothy O'Riordan's piece on public participation in risk management — but others are too diffuse. The emphasis is largely on infrequent hazards and accidents rather than long-term exposure to carcinogens and other toxins. Not all the essays are written from a social science perspective: there is a forthright defence of cost-benefit considerations by Christopher Foster, and a qualified endorsement of the importance of quantitative risk assessment by A. V. Cohen.

It is interesting to turn from this to some of the studies reported in Jan Gutteling and Oene Wiegman's book, *Exploring Risk Communication*. These seem to show that numerical information about risk often conveys no information to the recipient and may indeed be counter-productive. Gutteling and Wiegman provide a useful overview of the burgeoning field of risk communication, although there are few conclusions. They draw attention to the importance of fairness and mutual understanding, and quote K. E. Rowan, who said that risk communicators should abide by only one rule: 'be sincere'.

We have moved on from the time when communicating risk was rather like the stereotypical Englishman who tries to make foreigners understand by shouting louder rather than attempting to learn their language.

Or have we? One senses that Howard Margolis, author of *Dealing with Risk*, would not have much patience with social context and multidimensionality. He collapses all the dimensions of risk into a two-by-two matrix, with a yes/no for 'opportunity' or 'danger'. He argues that our response to risk is essentially visceral, and all other factors enter the situation as *post hoc* rationalizations. Even requirements of equity and justice are subsumed, almost by a sleight of hand, into utilitarian cost-benefit analysis — a serious error, as maximizing utility and removing inequity may be profoundly incompatible. According to Margolis, if our response does not agree with the expert consensus, we are simply victims of a cognitive illusion and, given the right circumstances, our perception can flip over into another cell of his two-by-two matrix.

There seems to be the germ of an interesting idea here, and some sensible suggestions: for example, that the overriding principle should be 'do no harm' when everything is taken into account (although showing that in practice would be complicated and contentious).

But any useful ideas that may be here struggle hard to emerge. Margolis may know something about risk perception, but he is way behind in the skills of risk communication.

The point about our being victims of cognitive illusions is illustrated with an intriguing example using poker chips, but this takes up far too much space. We do not reach a real example until more than halfway through the book, and then complex cases are disposed of in an extremely cursory fashion — there are only three pages on genetic engineering for example. More seriously, the author has imprisoned himself in a baroque style studded with terms that are never adequately defined, and some passages desperately need the attention of an editor, not to mention a proof reader.

The dust jacket of Margolis's book shows what looks like a chimney receding upwards from the observer. The colouring is blue, giving it a rather abstract, surreal appearance.

By contrast, the jacket of Peter Bernstein's *Against the Gods* shows Rembrandt's *Storm on the Sea of Galilee*, suggesting that we might be in for a tale of man pitted against the forces of nature, in which only divine intervention is certain to save us. Well, not really. It is more a case of man trying to survive the vagaries of the gambling tables and storms of the stock market with the help of mathematicians and economists.

The main theme of the book is a history of probability and statistics — Pascal's triangle, the normal distribution, Bayesian statistics, regression to the mean, chaos and game theory are all included. The narrative sweeps along, often breathlessly and sprinkled with superlatives: breakthroughs are "stunning", intellectual advances are "astonishing" and discoveries are "extraordinary". How much of the mathematics and statistics will be intelligible to the uninitiated is hard to say, and parts could certainly be made clearer. For example, although the normal distribution is referred to a great deal, there is no diagram of it.

But there are lots of good anecdotes, some of them amusing. My favourite is the one about the air force weather forecaster during the Second World War who realized that long-range forecasts were no better than numbers pulled out of a hat. On being told this, the commanding officer replied that he was "well aware that the forecasts are no good". However, he needed them "for planning purposes".

The historical journey with Bernstein is fast and exhilarating at times, but the landscape that we see seems to lack depth, with many more prices than values, and utility invariably in the foreground. In the end, though, we seem to have arrived at much the same viewpoint as that described by the rather maligned social scientists, albeit by a rather different route — despite the heroic efforts of man and computer, human beings do not always behave as economic theorists believe they should.

As John Maynard Keynes said: "Human decisions affecting the future, whether personal or political or economic, cannot depend upon strict mathematical expectation, since the bases for making such calculations do not exist."

So will there be a serious accident in the Channel Tunnel within, say, the next decade?

Personally I'll risk it. But I wouldn't bet on it. □

David Sumner is at Moss Park, Ravenstone, Whithorn, Dumfries and Galloway DG8 8DR, UK.

Foray beyond Fourier

Francois G. Meyer

The World According to Wavelets: The Story of a Mathematical Technique in the Making. By Barbara Burke Hubbard. A. K. Peters: 1996. Pp. 264. £26, \$38.

MANY natural phenomena can be interpreted as hierarchical superpositions of more elementary phenomena that occur at different scales or resolutions. The large-scale distribution of galaxies in the Universe can be described as a hierarchy of clusters and super-clusters, for instance. Another example is the study of turbulence, where different structures appear at very different scales.

In the physical sciences as well as in biology or medicine, many signals need to be analysed and interpreted at different resolutions. Indeed, singularities such as spikes, transients or abrupt changes are often more important than the general trend of the signal studied. Multiresolution analysis allows a signal to be described as the superposition of a very smooth average, which gives the trend, and a sequence of finer and finer details.

Until recently engineers and physicists did not have access to a mathematical theory for the multiresolution representation of functions. A function could only be studied either in its natural domain (space or time) or in the frequency domain using a Fourier transform. The Fourier transform is adequate for analysing signals whose statistical properties are stationary in time. But the Fourier representation is ill suited for other signals, such as seismic data, a speech signal or an electrocardiogram. Indeed it is difficult to characterize the local properties of a function in terms of the Fourier coefficients.

In the past ten years, a mathematical theory of multiresolution has been developed, and new building blocks called wavelets now make it possible to represent a signal in terms of small oscillating waveforms. Wavelets allow a signal to be observed at different resolutions, and its local behaviour to be analysed as well as its general trend.

As Barbara Burke Hubbard explains in her book, wavelet theory is a unification of ideas from different fields: engineering (image and speech compression), physics (coherent states, renormalization groups) and pure mathematics (Littlewood–Paley analysis, and study of Calderon–Zygmund operators).

The success of wavelets stems from the fact that they bring a new mathematical language to analyse and represent functions and signals. Wavelets lend themselves to efficient computational

algorithms. They are being used successfully in a wide variety of fields: signal processing, numerical analysis, statistics, computer vision and computer graphics.

This is not a textbook or review volume but a very original work that provides a self-contained introduction to the basic ideas of wavelets for scientists with little competence in mathematics. It is in two main parts. The first offers an accessible and enthralling introduction to wavelets including a review of Fourier analysis, a history of the various origins of wavelets and a description of wavelet techniques and their applications.

The reader travels a fascinating journey inside the world of mathematicians and physicists, and encounters the difficulties, frustrations and joys of research. The first part concludes with a general summary of wavelets and future directions of research in the field. The author keeps the style light, with a stress on intuition and clarity. There are many quotations from leading wavelet researchers.

The second part provides a self-contained mathematical description of Fourier analysis and wavelet analysis. This part emphasizes the comprehension of concepts rather than the simple acquisition of mathematical formulas. It can serve as an introduction for readers interested in using wavelets. It also offers the nonspecialist insight into the beauty and power of wavelet analysis. □

Francois G. Meyer is in the Department of Diagnostic Radiology, Yale University School of Medicine, 333 Cedar Street, PO Box 208042, New Haven, Connecticut 06520-8042, USA.

New in paperback

Soul Searching: Human Nature and Supernatural Belief by Nicholas Humphrey. Vintage, £6.99. "Humphrey is good at teasing out the contradictions and tensions in parapsychologists' claims. He also does a good job of underlining the curious (and inexplicable) banality of the supposed psi-phenomena", Tim Crane, *Nature* **379**, 685 (1996).

Inevitable Illusions: How Mistakes of Reason Rule Our Minds by Massimo Piattelli-Palmarini. Wiley, \$15.95, £12.99. "Delightful, informal... the best popular book in this field", Martin Gardner, *Nature* **374**, 25 (1995).

Entropy and the Magic Flute by Harold J. Morowitz. Oxford University Press, \$13.95. An experimental biologist and acclaimed essayist reflects on a diverse range of topics from litmus paper to the hippopotamus.

The Great Human Diasporas: The History of Diversity and Evolution by L. L. Cavalli-Sforza. Helix, \$15, £12.95. Reviewed by Robert Foley in *Nature* **377**, 493 (1996).

Magnetic fields in galaxies and beyond

Ellen G. Zweibel & Carl Heiles

Astronomical magnetic fields are generally strong enough to influence the dynamics of gas in present-day galaxies, and may have played an important role in the formation and early evolution of galaxies. Yet the origin of these magnetic fields remains controversial, and observational tests that could discriminate between competing theories will challenge the capabilities of telescopes now under construction.

PLANETS and stars, the tenuous interstellar gas in galaxies¹, and the even more tenuous medium in clusters of galaxies, are all magnetized². The epoch when these fields originated, and the mechanisms involved, are still uncertain. Both are aspects of what is arguably the most important issue in galactic magnetism: how the fields are generated and how they evolve. Observations of magnetic fields in galaxies and clusters of galaxies could in principle distinguish between theories in which the state of the field is nearly independent of initial conditions (dynamo theory) and theories in which the state depends on the properties of a pregalactic magnetic field (primordial theory). Neither theory is unequivocally supported by observation, and neither is complete.

Magnetic fields in galaxies are strong enough to significantly affect interstellar gas dynamics on both global scales characteristic of galactic structure³ and small scales characteristic of star formation^{4,5}. Cosmic rays, which heat, ionize and supply significant pressure to the interstellar medium are magnetically confined to galaxies and accelerated electromagnetically.

Magnetic fields could have affected galaxy formation and early evolution. Magnetic forces could have caused density fluctuations to form in primordial gas and influenced whether they collapsed to form bound objects^{6,7}. The main effects of galactic magnetic fields on star formation—removal of prestellar angular momentum and support of interstellar clouds against gravitational collapse— influence the distribution of stellar masses⁸. Without magnetic fields, this distribution might shift towards larger masses, as seems to occur early in galactic evolution⁸. Interstellar gas would also be colder and less stable against gravitational collapse into stars in

the absence of cosmic rays, which could not exist without magnetic fields.

There is evidence that magnetic fields were present in moderately young galaxies. The search could be extended, by observing radiation at longer wavelengths and with greater sensitivity, to even younger objects. As measurements of the cosmic background radiation are refined it may become possible to detect polarization produced by an intergalactic field that was present long before galaxies formed⁹. But at present, it is the nearby galaxies and galaxy clusters which are the most thoroughly observed.

Here we focus on observations and theories most relevant to the generation and maintenance of galactic magnetic fields. Although all the leading theories address certain aspects of the problem, none is yet complete. Much has been learned from the observations, but some important predictions of theories remain untested.

Magnetic fields in galaxies

External galaxies afford a 'bird's-eye view' of the global field structure that is unattainable in the Milky Way. In contrast, studies of the Milky Way provide spatial detail that is unattainable in other galaxies. In external galaxies, the orientation and degree of regularity of the field components in the plane of the sky, and the strength and regularity of the line-of-sight component, can be derived from the intensity, polarization and Faraday rotation of the diffuse synchrotron radiation. In the Milky Way, measurements of the line-of-sight component are also available through Faraday rotation measures (RMs) of radio sources and pulsars and Zeeman splitting measurements, while detailed optical and

BOX 1 Observational tracers of magnetic field

OBSERVATIONAL tracers respond only to either the plane-of-the-sky component $B_{\perp}$ or the line-of-sight component $B_{\parallel}$. The magnetic field has a uniform component B_0 and a random component B_r ; the total field $B_t^2 = B_0^2 + B_r^2$.

For $B_{\perp}$, we have:

(1) Synchrotron radiation intensity I_s . $I_s \propto N(E)B_{\perp}^2$, where $N(E)$ measures the density of relativistic electrons in the relevant energy range and x depends on the energy spectrum of the electrons; typically $x \approx 1.8$, so I_s responds roughly to $B_{\perp,t}^2$. In practice, $N(E)$ is usually not known so the strength of $B_{\perp}$ is indeterminate without assumptions, such as that of minimum total energy, in which case $I_s = KB_{\perp,t}^{7/2}$, where K is a constant which depends on various assumptions⁶⁷.

(2) Synchrotron radiation polarization. The direction of polarization is perpendicular to $B_{\perp}$ and the strength of polarization $\propto B_{\perp,u}^2/B_{\perp,t}^2$, indicating the randomness of the field⁶⁷.

(3) Polarization of starlight. The direction is parallel to $B_{\perp}$ and the strength of polarization depends on a number of geometrical and physical factors⁴.

(4) Polarization of infrared emission from dust. The direction is perpendicular to $B_{\perp}$ and the strength depends on a number of geometrical and physical factors⁶⁸.

For $B_{\parallel}$, we have:

(1) Faraday rotation. As a result of propagation through an ionized medium, the direction of linear polarization of a background radio source observed at wavelength λ rotates with λ^2 as measured by the rotation measure $RM \propto \int n_e B_{\parallel} ds$ (rad m⁻²), where n_e is the electron density and the integral is over the path length. Faraday rotation samples the warm ionized medium, which usually occupies a very small fraction of the interstellar volume; thus RMs may not provide a representative sample of the typical interstellar field. The direction of rotation indicates the sign of $B_{\parallel,u}$ but the strength of $B_{\parallel}$ is indeterminate without knowing the distribution of n_e (ref. 69).

(2) Zeeman splitting. Under normal circumstances in the interstellar medium, the Zeeman splitting of a spectral line is much smaller than the line width, and in this case the splitting is proportional to $B_{\parallel,u}^2$ rather than $B_{\parallel}$.

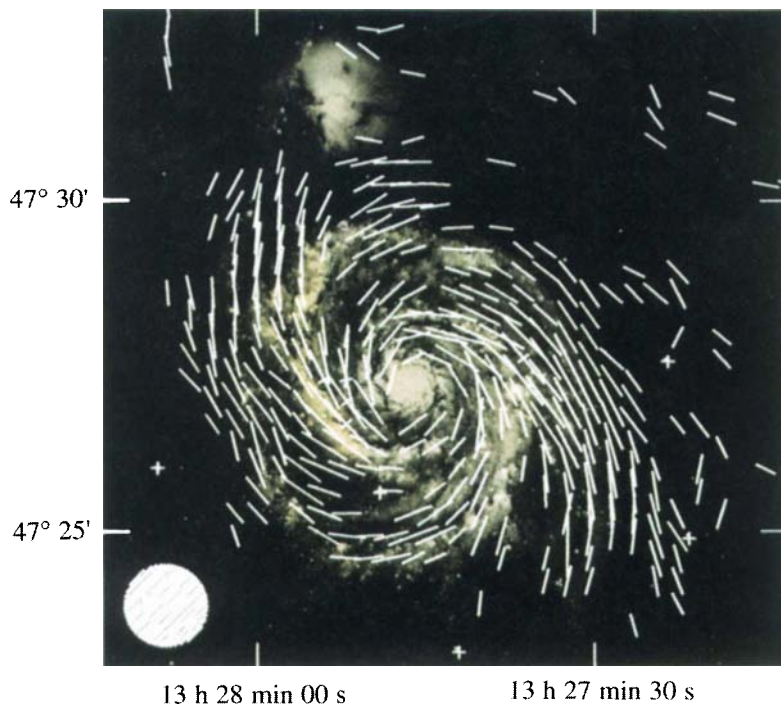


FIG. 1 Magnetic field vectors (white bars) for the galaxy M51⁷⁰, derived from the polarized continuum intensity at a wavelength of 2.8 cm, which is thought to be synchrotron radiation. The lengths of the vectors are proportional to intensity. The map is superimposed on an optical image. The white striped circle in the lower left corner indicates the size of the telescope beam. The crosses represent foreground stars.

polarization maps which provide the orientation of the field in the plane of the sky are also available (see Box 1). The field morphology and its degree of regularity constrain theories of the origin and evolution of galactic magnetic fields.

Observations of large numbers of external galaxies can reveal correlations between global galactic parameters and magnetic field structure. The best documented example is the correlation of 21-cm continuum intensity with far-infrared luminosity^{10,11}. The former depends on the relativistic electron population and magnetic energy density, and the latter is a measure of the rate at which massive stars are formed. The correlation holds over a wide range of spiral galaxy subtypes and star formation rates, and is not yet fully explained.

In external galaxies the field lines closely follow the spiral arms (Fig. 1). In some cases, the magnetic spirals are better defined than the material spiral arms: two outstanding examples are the galaxy NGC6946, in which the magnetic arms are particularly well defined in the interarm regions^{12,13}, and the flocculent galaxy NGC5055, which has ill-defined material arms but a definite spiral field¹³. A leading theory of spiral structure in galaxies, the density wave theory, predicts the inclination angle of the field lines to be larger inside spiral arms than in interarm regions¹⁴, but this is not observed¹³.

Galactic spiral fields are further subdivided into axisymmetric spirals (ASS) and bisymmetric spirals (BSS), which refer to even and odd parity, respectively, with respect to rotation by an angle π about the galactic centre¹ (Fig. 2). Present theories associate primordial fields with BSS configurations, and fields generated by dynamos with ASS configurations, but the distinction is not absolute, a matter we return to in the next section.

A minority of galaxies, which includes the Milky Way, have extensive radio haloes, indicating large scale heights for the magnetic field. It is unclear whether these halo fields originated in the disk, or independently of the disk fields.

It is standard practice to estimate total magnetic field strengths B_t in galaxies from synchrotron volume emissivity, assuming, in the absence of independent measurements of the relativistic particle density, that there is equipartition of energy between the magnetic field and relativistic particles. Equipartition appears to hold in the Milky Way^{15,16}, but field strengths in the starburst galaxy M82, and

in the Magellanic Clouds, exceed the equipartition field¹⁷. Whether equipartition holds in other galaxies has not yet been determined observationally, or justified by theory. With these caveats in mind, we note that average equipartition field strengths in galaxies range from 4 μ G for M33 to 12 μ G in NGC6946, and 19 μ G in NGC 2276^{18,19}, with the fields increasing toward the centres of the galaxies, and sometimes also within spiral arms.

The weight of the interstellar gas required to gravitationally confine magnetic fields to galaxies increases as B_t^2 , apart from factors that depend on field configuration. This should produce a correlation between magnetic field strength and hydrogen column density N_H and/or scale height of galactic disks. Indeed, the relationship $B_t^2 \propto N_H$ has recently been derived for a sample of 43 galaxies²⁰.

It is important for both observational and theoretical reasons to distinguish the uniform (or spatially averaged) field B_u from the total or r.m.s. field B_t . Measurements of the polarization p of synchrotron radiation at high radio frequencies, where Faraday effects are small, can be used to estimate the relative energies of the mean and random components of the field (see Box 1). In external galaxies the energy in the uniform field is about two-thirds of the energy of the random field except in regions of copious star formation, where it drops to as little as 5% (ref. 15). This is attributed to the dynamical effects of massive stars, which deliver momentum and energy to the interstellar gas through powerful winds and ultimately through supernova explosions. These phenomena produce bubble-like distortions of the field and generate hydromagnetic turbulence as the remnants cool and fragment²¹ and interact with interstellar clouds^{22,23}.

In the Milky Way we see the details; 'trees' instead of the 'forest'. Polarimetry of thousands of stars within a few kiloparsecs shows that B_t is parallel to the Galactic plane²⁴ (Fig. 3), while field strength and direction may be derived from the RMs of Galactic pulsars^{25,26} and of extragalactic radio sources²⁷. Once local features are properly removed, the mean field directions found by polarimetry and RM analysis are similar. The starlight data, which are superior on both statistical and systematic grounds, reveal²⁸ a spiral field inclined to the azimuth by $7.2^\circ \pm 4.1^\circ$, which probably differs significantly from the local spiral arm azimuthal inclination of 12.5° . The RM data show two reversals in field direction interior

to the solar Galactocentric orbit²⁶, both between spiral arms, as is also seen in the galaxy M81²⁹. There may also be two reversals in the outer galaxy²⁷. Thus, our Galaxy may have BSS.

Attempts to characterize the amplitude and typical scale length of magnetic field fluctuations have not yet produced a self-consistent picture³⁰. Polarization measurements of Galactic synchrotron radiation suggest that the energy in the B_u is slightly less than half the energy in fluctuations³¹. Modelling of pulsar RM data, on the other hand, suggest that less than 10% of the energy density is carried by the uniform component^{32,33}. Moreover, there is evidence for fluctuations in the field over distances ranging from less than a parsec (ref. 34) to 50–100 parsecs (refs 32, 33) to as large as a kiloparsec (ref. 35). Each type of measurement can, by its nature, sample only a small range of fluctuation sizes, and each reports a fluctuation amplitude at least as large as the amplitude of the uniform field. It is possible that these measurements, which sample a wide range of conditions in the interstellar medium, in some cases reflect only localized fluctuations which do not pervade the entire Galactic disk. Heiles¹⁶ has adopted $B_u = 2.2 \mu\text{G}$ and $B_u^2/B_i^2 \approx 0.27$ in the solar vicinity. This is a weighted mean of several results.

The study of individual magnetic-field structures and fluctuations leads to tests of magnetohydrodynamic theory applied to the interstellar medium. On scales of $\sim 100 \text{ pc}$ we see magnetized interstellar shells produced by individual and clustered supernovae. This is evident for nearby objects in the optical polarization of starlight, spectacularly so in the case of the North Polar Spur (Radio Loop I)³⁶, which is the huge ($\sim 120^\circ$ diameter) circular structure centred near Galactic longitude (l) -30° , latitude (b) 20° ; Fig. 3. Measurements of Zeeman splitting of the 21-cm line^{37,38} and of Faraday rotation³⁹ show that the magnetic field is enhanced by compression in these shells, as expected from theory⁴⁰.

At the $\sim 10\text{-pc}$ scale, starlight polarization maps show magnetic distortions near molecular clouds. The dense gas near the Orion nebula is associated with a highly perturbed field (Fig. 3), which appears to have been deformed by the multiple supernova explosions that produced the Eridanus Loop⁴¹. In contrast, the polarization map of the Taurus complex exhibits a large-scale wavy pattern whose physical significance remains obscure⁴² and which appears unrelated to the shapes of nearby molecular clouds. Thus, in contrast to shells (which are reasonably well understood), the nature of these smaller-scale features in molecular gas remains mysterious.

Magnetic field origin and maintenance

It is generally accepted that magnetic fields were amplified from insignificant cosmological values to their present galactic values by the inductive effect of fluid motions. The field energy can be increased in this way approximately up to equipartition with the kinetic energy in fluid turbulence, in rough agreement with what is observed in galaxies. However, there is a serious problem with this picture.

If there is no magnetic diffusion (that is, the magnetic Reynolds number $\text{Re}_m \rightarrow \infty$; see Box 2) and if the field occupies a constant volume then an increase in magnetic field strength must be accompanied by a proportional increase in field-line length, and thus progressive tangling of the field. As the ohmic resistivity of interstellar plasma is extremely low, one might expect the r.m.s. galactic field to exceed the mean field by many orders of magnitude, in striking contrast to their near equality.

This difficulty might be resolved in two ways, which are not mutually exclusive. First, interstellar turbulence could increase the magnetic diffusivity above the ohmic value by many orders of magnitude (an approximate analogy is that cream in coffee becomes smoothly distributed by diffusion, but it does so more quickly if stirred). Second, in the turbulent amplification picture the degree of cancellation between neighbouring field lines depends on details of the fluid flow. (This may be appreciated by representing a loop of magnetic field by an elastic band with

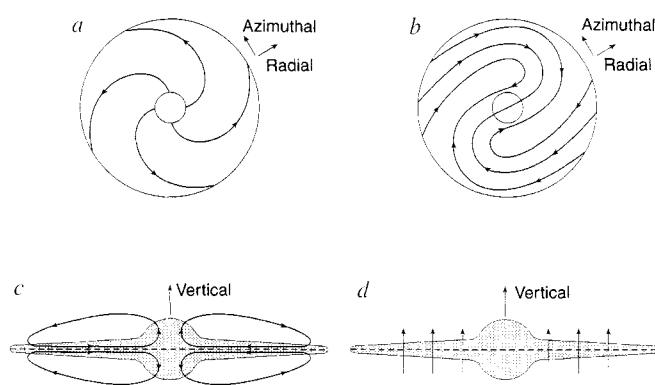


FIG. 2 Sketches of the magnetic field configurations in disk galaxies according to various models. The bold lines are magnetic field lines, with the field direction indicated by arrows. *a*, Face-on view of axisymmetric spiral structure (ASS). The pattern must change at the centre of the galaxy. *b*, Face-on view of bisymmetric spiral structure (BSS). *c*, Edge-on view of quadrupolar (even-parity) magnetic structure. *d*, Edge-on view of odd-parity magnetic structure; sometimes called dipolar structure, although the field lines need not all return to the galaxy, as they would for a standard magnetic dipole.

arrows drawn along it. Folding the band over on itself demonstrates cancellation; twisting it once before folding it represents reinforcement). The magnetic forces which act on the fluid (Box (2) equation (6)) might suppress those motions which lead to a high degree of cancellation so that it is possible to generate a field which is smooth on large scales⁴³.

The two leading pictures for the origin and maintenance of galactic magnetic fields—the dynamo and primordial pictures—differ by their treatment of the foregoing issues. In dynamo theory, the magnetic diffusivity of galaxies is high and the field would quickly decay if it were not perpetually regenerated by fluid motions. In primordial theory the magnetic diffusivity of galaxies is low. Magnetic flux is neither created nor destroyed in large amounts in galaxies at present, and the existing field was created at an earlier epoch.

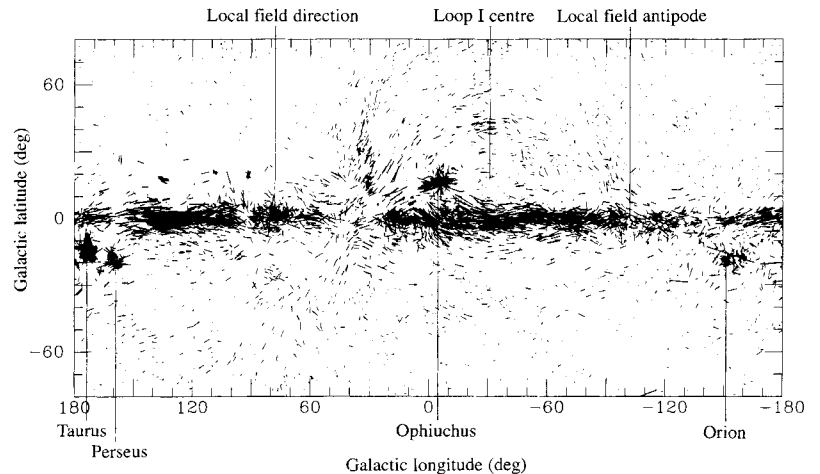
Dynamo theory. Galactic dynamo theory predicts the mean interstellar magnetic field B_u under the action of large-scale flows (for example, galactic rotation or a galactic wind) and small-scale turbulent flows. The range of scales involved is too large to model directly, so the effects of turbulence are parametrized by transport coefficients representing both induction and destruction of the mean field.

In the simplest versions of mean field dynamo theory, the governing equations are linear in B_u , and the field can grow exponentially if the resistivity is low enough (but non-zero). The galaxy model and boundary conditions adopted lead to a discrete set of modes, and the fastest growing one is usually assumed to dominate. Many simulations go beyond linear theory by attempting to take into account the forces exerted by B_u , which inhibit dynamo action. This ‘quenching’ is assumed to become significant when the energy in the mean field comes into equipartition with the turbulent energy, so B_u saturates at about its equipartition value⁴⁴, in approximate agreement with observations.

Most galactic disk dynamos generate fields which are predominantly azimuthal—the result of strong shear by differential rotation—and have quadrupolar symmetry about the galactic plane: the vertical component B_z reverses across the midplane whereas the azimuthal and radial components B_{ϕ} and B_r do not¹ (Fig. 2). At small galactocentric radii, solid body rotation leads to predominantly dipolar fields.

The fields produced by these dynamos are generally ASS rather than BSS. This result may change if the transport is directionally anisotropic, as found from calculations of transport for specific

FIG. 3 Starlight polarization based on a catalogue of 7,503 stars⁷¹ and several additional studies concentrating on the named star-forming regions (A. A. Goodman, personal communication; available from agoodman@cfa.harvard.edu). The orientation of each line is along the polarization, which is parallel to $B_{\perp}$, the projection of B_z on the plane of the sky. The length of the line indicates the percentage polarization p , for which there are two scales: thin lines have $p < 0.6\%$ and are proportionately four times longer than thick lines. The local field points towards Galactic longitude $l \approx 80^\circ$ (and away from -100°). Away from these directions, the stellar polarizations are oriented along $B_{\perp}$ and are systematically parallel to the Galactic plane. In contrast, towards these directions $B_{\perp}$ 'recedes into the distance' and the orientations are ill-defined. Apart from this very systematic behaviour there also exist obvious perturbations. The field lines have been swept into the edge of the 120° -diameter spherical shell of Loop I, formed by repeated shocks from supernovae exploding in the Scorpius/Ophiuchus star clusters. The field lines in the active star-forming regions Taurus, Perseus, Ophiuchus and Orion are perturbed on small scales and depart from the large-scale pattern.



interstellar turbulence models^{45,46}. Unpublished calculations by P. Fox and K. Ferriere show that $B_{u\phi}$ can reverse with radius in such models even when the field is axisymmetric. This may provide an alternative to BSS for explaining the field reversals seen in the Milky Way.

Although the dynamo equations are typically solved only for B_u , they can also be used to compute B_r . It has been argued⁴⁷ that under interstellar conditions B_r grows much faster than B_u , as expected for a field which grows in a nondiffusive medium and exerts negligible forces⁴⁸.

Primordial theory. These theories⁴⁹ postulate that galactic disks are endowed with large-scale magnetic fields of microgauss strength at the time they form. Resistivity may be ignored, so the field evolves only throughout its advection by flow. Differential rotation, which may be a function of both radius and vertical height, wraps the field into a bisymmetric spiral. None of the field components reverses across the galactic plane, unless the field was initially nearly vertical, in which case $B_{u\phi}$ would be an odd function of z (ref. 50). Differential rotation makes the azimuthal component $B_{u\phi}$ predominate in magnitude and reverse in direction with radius. The magnitude of $B_{u\phi}$ increases linearly with time t , and the distance between reversals decreases as t^{-1} .

The number of reversals observed in the Milky Way is about one-tenth of the number predicted by the model. Inclusion of vertical transport of field lines in the model reduces the number of predicted reversals. Such transport could result from a 'galactic fountain'⁵¹ of hot upflow and cool downflow⁵², or from the ambipolar drift of ions relative to neutral particles⁵³ down the magnetic pressure gradient and away from the galactic plane⁵⁴.

The approximate equipartition of B_u with turbulent pressure is unexplained by the model, but might arise through the buoyancy instabilities which operate if B_r becomes too large⁵⁵, and could maintain the field strength near marginal stability.

Interstellar turbulence does not play an important role in primordial field models. Theory predicts that the random field and fluid turbulence are roughly in equipartition, consistent with a statistically steady state⁵⁶ in which there is no net induction or destruction of magnetic field.

Evidence for dynamo and primordial theories. The mean field dynamo and primordial theories predict opposite magnetic parities of the vertical field with respect to the galactic plane, suggesting an important observational test which has not yet been applied to any galaxy. The observations are difficult because B_w is evidently small in most galaxies, and the vertical field of fluctuating sign which is naturally produced by superbubbles⁵⁰ must be subtracted away.

A second test, the parity of the field with respect to rotations by π about the galactic centre, is more provisional both observationally and theoretically. The primordial theory predicts odd parity and reversals with radius (a bisymmetric spiral field), whereas most current dynamo models predict even parity and no reversals. Most galaxies exhibit no recognizable large-scale pattern. Reversals are seen in the Milky Way and in M81 but not in M31 and IC342; observing more galaxies at higher angular resolution might reveal additional cases of reversal. The predictions of dynamo theory might change with more realistic parametrization of the effects of small-scale turbulence^{45,46}.

The reversals of $B_{u\phi}$ seen in the Milky Way and in M81 occur between spiral arms^{26,27,29}. Although the number of galaxies in which reversals are seen is so small that the significance of this is uncertain, it is worth mentioning that neither dynamo nor primordial theory predicts any relationship between magnetic spiral structure and stellar/hydrodynamic spiral structure. The latter is only beginning to be included in dynamo models¹, and is not included in primordial models.

A final test of the theories entails understanding magnetic field diffusion in galaxies. It has not yet been demonstrated that a large-scale field of significant magnitude can be generated by a dynamo without magnetic reconnection (topology-destroying merging of the field), at a rate which is virtually independent of ohmic resistivity⁵⁵. It is uncertain whether this rapid rate can be established in galaxies. The viability of both galactic dynamo and primordial field theories hinge to a remarkable extent on the nature of magnetic reconnection under interstellar conditions.

Fields at the largest scales and earliest times

The discovery of a pervasive intergalactic field would be very significant. It would suggest magnetogenesis before galaxy formation, and show that the formation of galaxies^{6,7} and the first stars⁸ could have been affected by magnetic fields. Similarly, detection of magnetic fields in objects which are at cosmological distances, and therefore are viewed in their distant past, holds important clues to the origin of galactic magnetic fields.

Intergalactic fields. Although searches for magnetic fields throughout the intergalactic medium have led only to upper limits⁵⁶, detections of synchrotron radiation and Faraday rotation from clusters of galaxies provide secure evidence for magnetic fields in the intracluster medium. As this gas contains heavy elements, at least part of it is thought to have cycled through stars and been ejected from galaxies, accompanied by its magnetic field. Therefore, the implications for pregalactic magnetogenesis

are uncertain. However, the cluster magnetic fields appear to be stronger than the randomly oriented field of a few hundredths of a microgauss that would be the outcome of galactic origin. This may be evidence for a dynamo operating now and/or in the past in clusters of galaxies. If the field were amplified up to equipartition with the kinetic energy the field strength would be about $1 \mu\text{G}$, and would tend to have random sign.

Possible evidence for a coherent field follows from the excess mean RMs (above those of a control sample) of extragalactic radio sources located behind clusters of galaxies⁵⁷. The excess RM, typically $\sim 30 \text{ rad m}^{-2}$, with typical electron densities (derived from the thermal X-ray emission) and cluster diameters requires $B_0 \approx 0.02 \mu\text{G}$. However, the excess RM could also be produced in a smaller region where field strengths are higher.

The polarization of the synchrotron radiation from clusters is very low, suggesting that the field is almost entirely random. The intensity of relativistic electrons in one cluster, the Coma cluster, is bounded above by upper limits on γ -ray bremsstrahlung⁵⁸ and X-ray inverse Compton photons⁵⁹ from the cluster. This leads to lower limits on the total field strength: $B_t > 0.4 \mu\text{G}$ (ref. 58) and $B_t > 0.1 \mu\text{G}$ (ref. 59), respectively. Less direct methods of estimating the field have led to larger values, of the order of $1 \mu\text{G}$, but these values depend on the details of the models and we consider them less reliable.

Whether the field is one or a few tenths of a microgauss, it seems to be one to two orders of magnitude larger than the field that would be obtained from galactic ejecta alone. It appears that dynamos operate, or once operated, in clusters of galaxies.

Magnetic fields in the distant past. It is important to determine whether magnetic fields exist in very young galaxies, not only because of the implications for magnetogenesis but also because of the role of magnetic fields in star formation⁸.

The first studies of Faraday rotation of quasars lying behind high-redshift galaxies concluded that the absorption line systems did indeed produce statistically significant RM detections⁶⁰ and also found that damped Lyman- α systems produce significantly higher RMs than undamped systems⁶¹ (the former are thought to be galactic disks; the latter may be galactic haloes). Reconsideration of the systematic errors later called the statistical significance of this result in question^{62,63}. Nevertheless, we feel that future studies will improve the statistics and regain the original conclusions, and in particular that the excess RMs associated with damped compared to undamped systems will be upheld. Also, a radio interferometric study of the remarkable damped system in front of PKS1229 – 021 measured RMs that change sign on a length scale that is similar to those seen for magnetic reversals in some nearby spiral galaxies⁶⁴ and the inferred magnetic field strength is a few microgauss.

The existence of microgauss-strength magnetic fields in disk galaxies at redshift $z \approx 1$ is a challenge to dynamo theory. Such galaxies should have rotated at most about ten times. As the rotation rate is thought, from dynamo theory, to be an upper bound on the mean magnetic field amplification rate, and 'seed' magnetic fields generated by charge separation in the early Universe are estimated to be at least 15 orders of magnitude smaller than galactic fields at the present epoch⁶⁵, it seems unlikely that straightforward application of mean field galactic dynamo theory can explain the rotation measures seen at moderate redshift. It has been proposed instead that the seed fields are first amplified by a small-scale turbulent dynamo, and that a mean field is then created from the turbulent field by a conventional mean field galactic dynamo¹. A recent model invokes strong turbulent amplification during galaxy formation⁶⁵.

The problems discussed above raise many of the same issues we encountered in describing galactic magnetic fields, but in a more extreme form. We are faced with evidence that highly conducting, turbulent plasmas with enormous diffusion times are pervaded by strong, possibly rather coherent magnetic fields which have grown in cosmologically short times. This great theoretical challenge needs to be constrained by a larger base of observations.

BOX 2 Magnetohydrodynamics

MAGNETOHYDRODYNAMICS (MHD) describes phenomena in magnetized gases which take place on length scales and timescales which are longer and slower than the Debye length and plasma period (which characterize charge separation effects) or the scales characterizing particle gyration about magnetic field lines. The non-relativistic version holds when the characteristic fluid velocities v are much less than c , the speed of light. Under these conditions, one can use a reduced form of Maxwell's equations for the electromagnetic fields $\mathbf{E}$ and $\mathbf{B}$ and the current density $\mathbf{J}$, which in gaussian c.g.s. units are

$$\nabla \cdot \mathbf{B} = 0 \quad (1)$$

$$\frac{\partial \mathbf{B}}{\partial t} = -c \nabla \times \mathbf{E} \quad (2)$$

$$\nabla \times \mathbf{B} = \frac{4\pi \mathbf{J}}{c} \quad (3)$$

Combining equation (2) with Ohm's law assuming resistivity η

$$\mathbf{E} + \frac{\mathbf{v}}{c} \times \mathbf{B} = \eta \mathbf{J} \quad (4)$$

yields the magnetic induction equation

$$\frac{\partial \mathbf{B}}{\partial t} = \nabla \times (\mathbf{v} \times \mathbf{B}) + \frac{\eta c^2}{4\pi} \nabla^2 \mathbf{B} \quad (5)$$

where we have assumed that the magnetic diffusivity $\eta c^2/4\pi$ is independent of position. When $\eta = 0$, it can be shown that the magnetic field is frozen to the fluid in the sense that the magnetic flux $\Phi \equiv \int \mathbf{B} \cdot d\mathbf{S}$ through any co-moving surface of fluid is constant. When η is non-zero, the importance of resistivity in systems with characteristic velocities v and lengthscales L is measured by the ratio of the resistive decay time $4\pi L^2/\eta c^2$ to the dynamical time L/v . This ratio, $4\pi L v/\eta c^2$, is known as the magnetic Reynolds number, Re_m .

The electromagnetic force density $\mathbf{F}_M$ in MHD is given by

$$\mathbf{F}_M = \frac{\mathbf{J} \times \mathbf{B}}{c} = \frac{(\nabla \times \mathbf{B}) \times \mathbf{B}}{4\pi} \quad (6)$$

which can also be written as

$$\mathbf{F}_M = -\nabla \frac{B^2}{8\pi} + \frac{\mathbf{B} \cdot \nabla \mathbf{B}}{4\pi} \quad (7)$$

The first term in equation (7) represents magnetic pressure, and the second term is magnetic tension. The latter term gives rise to tension waves which propagate at the characteristic velocity known as the Alfvén speed

$$v_A \equiv \frac{B}{(4\pi\rho)^{1/2}}$$

where ρ is the mass density.

Future directions

Perhaps the most important outstanding issue in large-scale magnetism relates to the origin and maintenance of galactic magnetic fields: are they the dynamo-amplified versions of a tiny initial seed field, or are they the compressed and wound-up versions of a much stronger pre-existing intergalactic field?

Mean field dynamo theory predicts that the magnetic field has quadrupolar structure (B_z antisymmetric and B_r , B_θ symmetric about the galactic plane; Fig. 2). The symmetries in the plane are somewhat less certain, but axisymmetric spirals are generally favoured over bisymmetric spirals by the present models. It is assumed in most galactic dynamo theories that the magnetic diffusivity of galaxies is many orders of magnitude larger than the ohmic value because the field is rapidly mixed by turbulence, but this has been questioned⁶⁶. A better understanding of turbulent mixing will probably be achieved only by a combination of magnetohydrodynamic turbulence theory and numerical experiment.

The primordial theory predicts that B_z , B_r and usually B_θ are symmetric about the midplane. In the plane, there are reversals

with galactic radius and the symmetry is bisymmetric. The primordial theory requires fairly strong ($\sim 10^{-3}$ μ G) pregalactic magnetic fields, and must explain the approximate equality between the mean and random fields.

These predictions provide two observational tests of dynamo against primordial theory. One is the parity of the vertical field. The other is the symmetry of the field in the galactic plane: bisymmetric versus axisymmetric structure. The first of these tests would be more decisive. However, as B_z in the Milky Way is evidently quite small, this important test is difficult to carry out even in our own Galaxy. The second test has been more fruitful, with 3 out of 20 galaxies exhibiting recognizable symmetry¹ (one BSS, two ASS) and the Milky Way exhibiting reversals with radius, which may imply BSS. One might consider the score even, but the disappointing fact is that 80% of galaxies have unrecognizable symmetry. These symmetries are best revealed by observations of synchrotron polarization, Faraday rotation and Zeeman splitting, which are very difficult because of limited sensitivity, instrumental effects and interpretive ambiguities. Observers will have to push the existing radioastronomical facilities, together with the upgraded Arecibo and new Green Bank telescopes, to their limits.

The relationship between magnetic and other forms of galactic spiral structure, although known in only a few galaxies, is unexplained theoretically, and remains intriguing. A model of magnetic field interaction with galactic spiral structure developed some time ago¹⁴ predicts that the fields should be closely aligned with spiral arms following compressive spiral shocks, and that the inclination angle should decrease between spiral arms. Although overall the field lines tend to be aligned with spiral arms, the

predicted change of inclination angle has not been confirmed, possibly because of viewing effects¹⁵. A recent discussion of a new type of spiral density wave in which magnetic arms lie between spiral arms⁷² deserves further scrutiny in light of the recent discovery of such systems^{12,13}.

Whatever the origin of the fields, their properties are linked with the dynamics and energetics of interstellar gas. Magnetic fields and cosmic rays contribute significant vertical support to the gas³. A geometrically thick synchrotron halo is present in the Milky Way, but is one of a very few known haloes. Star formation is coupled to magnetic fields in observable ways, but other important aspects of coupling between interstellar gas and magnetic fields are not usually obvious from observations. For example, magnetic pressure reduces the volumes of supernova remnants and superbubbles, but shock waves driven by supernovae and superbubbles are the main driver of interstellar turbulence, without which there could be no galactic dynamo.

Taking these things together, we expect that magnetic fields play an important role in shaping the global structure of interstellar gas, while the state of the gas feeds back on the field. The nature of this self-consistent state is not yet observationally characterized or theoretically understood, and developing and testing models for it will be a great challenge in the future. $\square$

E.G.Z. is at the Joint Institute for Laboratory Astrophysics, Department of APAS, Campus Box 391, University of Colorado, Boulder, Colorado 80309-0440, USA. C.H. is at the Department of Astronomy, University of California, Berkeley, California 94720, USA.

1. Beck, R., Brandenburg, A., Moss, D., Shukurov, A. & Sokoloff, D. *Annu. Rev. Astron. Astrophys.* (submitted).
2. Kronberg, P. P. *Rep. Prog. Phys.* **57**, 325–382 (1994).
3. Boulares, A. & Cox, D. P. *Astrophys. J.* **365**, 544–558 (1990).
4. Heiles, C., Goodman, A. A., McKee, C. F. & Zweibel, E. G. in *Protostars and Planets III* (eds Levy, E. H. & Lunine, J. I.) 279–326 (Univ. Arizona Press, Tucson, 1993).
5. McKee, C. F., Zweibel, E. G., Heiles, C. & Goodman, A. A. in *Protostars and Planets III* (eds Levy, E. H. & Lunine, J. I.) 327–366 (Univ. Arizona Press, Tucson, 1993).
6. Wasserman, I. *Astrophys. J.* **224**, 337–343 (1978).
7. Kim, E.-J., Olinto, A. V. & Rosner, R. *Astrophys. J.* **468**, 28–50 (1996).
8. Rees, M. J. Q. J. R. *Astron. Soc. Trans.* **28**, 197–206 (1987).
9. Kosowsky, A. & Loeb, A. *Astrophys. J.* **469**, 1–7 (1996).
10. Dickey, J. M. & Salpeter, E. E. *Astrophys. J.* **284**, 461–470 (1984).
11. Garwood, R. W., Helou, G. & Dickey, J. M. *Astrophys. J.* **322**, 88–100 (1987).
12. Beck, R. & Hoernes, P. *Nature* **379**, 47–49 (1996).
13. Beck, R. in *Polarimetry of the Interstellar Medium* (eds Roberge, W. R. & Whittet, D. C. B.) 475–485 (Conf. Ser. Vol. 97, Astron. Soc. Pacific, San Francisco, 1996).
14. Roberts, W. W. & Yuan, C. *Astrophys. J.* **101**, 877–902 (1970).
15. Heiles, C. in *The Physics of the Interstellar Medium and Intergalactic Medium* (eds Ferrara, A., McKee, C. F., Heiles, C. & Shapiro, P. R.) 507–523 (Conf. Ser. Vol. 80, Astron. Soc. Pacific, San Francisco, 1995).
16. Heiles, C. in *Polarimetry of the Interstellar Medium* (eds Roberge, W. R. & Whittet, D. C. B.) 447–474 (Conf. Ser. Vol. 97, Astron. Soc. Pacific, San Francisco, 1996).
17. Chi, X. & Wolfendale, A. W. *Nature* **362**, 610–611 (1993).
18. Buczkowski, V. R. & Beck, R. *Astron. Astrophys.* **241**, 47–46 (1991).
19. Hummel, E. & Beck, R. *Astron. Astrophys.* **303**, 691–704 (1995).
20. Niklas, S. & Beck, R. *Astron. Astrophys.* (in the press).
21. McCray, R. & Kafatos, M. *Astrophys. J.* **317**, 190–196 (1987).
22. Spitzer, L. *Astrophys. J.* **262**, 315–321 (1982).
23. Miesch, M. S. & Zweibel, E. G. *Astrophys. J.* **432**, 622–640 (1994).
24. Mathewson, D. & Ford, V. L. *Mem. R. Astron. Soc.* **74**, 139–182 (1970).
25. Manchester, R. N. *Astrophys. J.* **188**, 637–644 (1974).
26. Rand, R. & Lyne, A. G. *Mon. Not. R. Astron. Soc.* **268**, 497–505 (1994).
27. Clegg, A. W., Cordes, J. M., Simonetti, J. H. & Kulkarni, S. R. *Astrophys. J.* **386**, 143–157 (1992).
28. Heiles, C. *Astrophys. J.* **462**, 316–325 (1996).
29. Krause, M., Beck, R. & Hummel, E. *Astron. Astrophys.* **217**, 17–30 (1989).
30. Zweibel, E. G. in *Polarimetry of the Interstellar Medium* (eds Roberge, W. G. & Whittet, D. C.) 486–503 (Conf. Ser. Vol. 97, Astron. Soc. Pacific, San Francisco, 1996).
31. Spoelstra, T. A. T. *Astron. Astrophys.* **135**, 238–248 (1984).
32. Rand, R. J. & Kulkarni, S. R. *Astrophys. J.* **343**, 760–772 (1989).
33. Ohno, H. & Shibata, S. *Mon. Not. R. Astron. Soc.* **262**, 953–962 (1993).
34. Minter, A. H. & Spangler, S. R. *Astrophys. J.* **458**, 194–214 (1996).
35. Jones, T. J., Klee, D. & Dickey, J. M. *Astrophys. J.* **389**, 602–615 (1992).
36. Egger, R. J. in *The Physics of the Interstellar Medium and Intergalactic Medium* (eds Ferrara, A., McKee, C. F., Heiles, C. & Shapiro, P. R.) 45–54 (Conf. Ser. Vol. 80, Astron. Soc. Pacific, San Francisco, 1995).
37. Heiles, C. *Astrophys. J.* **336**, 808–821 (1989).
38. Myers, P. C., Goodman, A. A., Gusten, R. & Heiles, C. *Astrophys. J.* **442**, 177–185 (1995).
39. Vallée, J. P. *Astrophys. J.* **419**, 670–673 (1993).
40. Ferrière, K. M., MacLow, M. M. & Zweibel, E. G. *Astrophys. J.* **375**, 239–253 (1991).
41. Heiles, C. *Astrophys. J.* (submitted).
42. Goodman, A. A., Bastien, P., Myers, P. C. & Ménard, F. *Astrophys. J.* **359**, 363–377 (1990).
43. Pouquet, A., Frisch, U. & Léorat, J. *J. Fluid Mech.* **77**, 321–354 (1976).
44. Moss, D., Brandenburg, A., Donner, K. J. & Tomasson, M. *Astrophys. J.* **409**, 179–189 (1993).
45. Ferrière, K. M. *Astrophys. J.* **404**, 162–184 (1993).
46. Ferrière, K. M. *Astrophys. J.* **409**, 248–261 (1993).
47. Kulsrud, R. M. & Anderson, S. W. *Astrophys. J.* **396**, 606–630 (1992).
48. Childress, S. & Gilbert, A. D. *Stretch, Twist, Fold: The Fast Dynamo* (Springer, Berlin, 1995).
49. Kulsrud, R. M. in *Proc. Int. Course and Workshop on Plasma Astrophysics* 531–535 (ESA, Noordwijk, 1986).
50. Zweibel, E. G. in *Physics of the Interstellar Medium and Intergalactic Medium* (eds Ferrara, A., McKee, C. F., Heiles, C. & Shapiro, P. R.) 524–544 (Conf. Ser. Vol. 80, Astron. Soc. Pacific, San Francisco, 1995).
51. Shapiro, P. R. & Field, G. B. *Astrophys. J.* **205**, 762–765 (1976).
52. Brandenburg, A., Moss, D. & Shukurov, A. *Mon. Not. R. Astron. Soc.* **276**, 651–662 (1995).
53. Howard, A. & Kulsrud, R. *Astrophys. J.* (submitted).
54. Parker, E. N. *Cosmical Magnetic Fields: Origin and Activity* (Oxford Univ. Press, 1979).
55. McKee, C. F. & Zweibel, E. G. *Astrophys. J.* **439**, 779–792 (1995).
56. Kronberg, P. P. & Simard-Normandin, M. *Nature* **263**, 653–656 (1976).
57. Kim, K.-T., Tribble, P. C. & Kronberg, P. P. *Astrophys. J.* **379**, 80–88 (1991).
58. Rephaeli, Y., Ulmer, M. & Gruber, D. *Astrophys. J.* **429**, 554–556 (1994).
59. Sreekumar, P. et al. *Astrophys. J.* **464**, 629–640 (1996).
60. Welter, G. L., Perry, J. J. & Kronberg, P. P. *Astrophys. J.* **279**, 19–29 (1984).
61. Wolfe, A. M., Lanzetta, K. M. & Oren, A. L. *Astrophys. J.* **388**, 17–22 (1992).
62. Perry, J. J., Watson, A. M. & Kronberg, P. P. *Astrophys. J.* **406**, 407–419 (1993).
63. Oren, A. L. & Wolfe, A. M. *Astrophys. J.* **445**, 614–641 (1995).
64. Kronberg, P. P., Perry, J. J. & Zukowski, E. L. H. *Astrophys. J.* **355**, L31–L34 (1992).
65. Kulsrud, R. M., Ryu, O., Cen, R. & Ostriker, J. P. *Astrophys. J.* (submitted).
66. Cattaneo, F. & Vainshtein, S. I. *Astrophys. J.* **376**, L21–L24 (1991).
67. Ginzburg, V. L. & Syrovatskii, S. I. *Annu. Rev. Astron. Astrophys.* **3**, 297–350 (1965).
68. Hildebrand, R. H. & Dragovan, M. *Astrophys. J.* **450**, 663–666 (1995).
69. Burn, B. J. *Mon. Not. R. Astron. Soc.* **133**, 67–83 (1965).
70. Neiningen, N. *Astron. Astrophys.* **263**, 30–36 (1992).
71. Mathewson, D. S., Ford, V. L., Kline, G., Neckel, Th. & Kranter, J. *Bull. Inform. CDS* **14**, 115; available at <http://edsweb.u-strasbg.fr/>
72. Fan, Z. & Lou, Y.-Q. *Nature*, **383**, 800–801 (1996).

ACKNOWLEDGEMENTS. We thank our colleagues for information and advice; we especially thank R. Beck, E. Berkhuisen, E. Blackman, A. Brandenburg, F. Cattaneo, G. Field, R. Kulsrud, C. McKee, S. Niklas, R. Rand, R. Rosner and M. Shull. R. Beck supplied the print from which Figure 1 was made. This work was supported by the US NSF and NASA.

CORRESPONDENCE should be addressed to E.G.Z. (e-mail: zweibel@lepton.colorado.edu).

A high deuterium abundance in the early Universe

Antoinette Songaila*, E. Joseph Wampler†
& Lennox L. Cowie*

* Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive,
Honolulu, Hawaii 96822, USA

† National Astronomical Observatory, Osawa, Mitaka, Tokyo 181, Japan

INTERGALACTIC gas clouds at high redshifts have element abundances that are close to primordial. The ratio of deuterium to hydrogen (D/H) within such clouds—which is determined from absorption lines in the spectra of more distant quasars that lie along the same line of sight—provides the best estimate of the density of baryons (Ω_b) in the Universe. Previous estimates of D/H in the early Universe have yielded values that differ by about an order of magnitude^{1–7}, with the lower values^{6,7} implying a high density of baryons that may be difficult to reconcile with both estimates of the primordial abundances of other light elements (especially ⁴He) and the known number of light neutrinos^{8–10}. The accuracy of such D/H determinations is heavily dependent on the inferred column density of neutral hydrogen in the absorbing clouds. Here we report an independent measurement of the neutral hydrogen column density in the cloud towards the quasar Q1937 – 1009, for which one of the low D/H values was derived⁶. Our measurement requires a substantial revision to the D/H value reported previously; we obtain a lower limit of $D/H > 4 \times 10^{-5}$ for this cloud, which implies $\Omega_b < 0.016$ for a Hubble constant of $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$. This reduced upper limit for the baryon density relieves any conflict with standard Big Bang nucleosynthesis⁶.

The accuracy of the measurements of D/H in high-redshift quasar absorption line clouds is heavily dependent on the neutral hydrogen column density of the absorbing cloud; different problems arise in different column density regimes. At sufficiently low column density^{1–5}, absorption lines of hydrogen high in the Lyman series become unsaturated and permit an accurate determination of H, but the corresponding weakness of deuterium increases the likelihood of significant contamination by a chance coincidence of weak foreground H I absorption. Alternatively, choosing a higher column density absorber minimizes such deuterium contamination but the saturation of the Lyman series increases the chance of wrongly estimating the H I column density. The two low measurements of D/H ($\sim 2.3 \times 10^{-5}$) towards Q1937 – 1009 (ref. 6) and Q1009 + 2956 (ref. 7) belong in this category.

In the latter case, the total amount of neutral hydrogen is hard to estimate because even the lines high in the Lyman series are nearly opaque. For the same reason, the exact velocity distribution of H I and D I absorption cannot be determined directly but must be plausibly modelled from the distribution of the weaker absorption of other ions and this procedure can be problematical¹¹. Towards Q1937 – 1009, Tytler *et al.*⁶ made the minimal assumption that the simplest two-component model that accounted for low ionization absorption was sufficient to account for H I and D I absorption. Recent investigation¹¹ has shown that alternative models of the absorption in this cloud that have as much as three times lower H I column density can be found, which would increase D/H to 7×10^{-5} , well outside the published⁶ errors.

Regardless of its actual velocity distribution, the total H I column density associated with a cloud can be found by observing the optical depth of the corresponding redshifted Lyman continuum. The different cloud models proposed^{6,11} to account for the absorption towards Q1937 – 1009 make different predictions for the amount of residual flux below the Lyman break; these predictions are difficult to distinguish at the required level with the

existing⁶ high-resolution echelle spectra obtained using the Keck HIRES spectrograph. In the violet spectral region, the orders of this spectrograph are closely spaced and the necessarily short slits limit the amount of sky available for sky subtraction. We have addressed this difficulty by using long-slit, lower-resolution spectroscopy to measure precisely the residual flux below the Lyman continuum break.

The measurements were obtained in August 1996 with the Low Resolution Imaging Spectrograph (LRIS) on the KeckI 10-m telescope on Mauna Kea, Hawaii and consist of a total of 45 minutes exposure at lower resolution (resolving power, $R \approx 300$) through a very wide slit (1.5 arcsec) and a higher-resolution ($R \approx 1,500$) exposure (40 minutes total) taken through a 0.7-arcsec slit. (Seeing was less than 0.8 arcsec for both observations.) The spectra were flux-calibrated with observations of a white dwarf star taken immediately after the target observations and at nearly identical zenith angle and declination, and wavelength-calibrated with observations of a krypton-mercury lamp in the same configuration. The two-dimensional spectral image of the quasar (Fig. 1) clearly shows a residual flux below the continuum break.

The flux-calibrated and sky-subtracted low-resolution spectrum is shown in Fig. 2a. Both the quasar flux level and the level of the unabsorbed continuum are needed for calculating the optical depth of the Lyman continuum. For Q1937 – 1009 the quasar's emission redshift is $z = 3.805$, only slightly higher than the redshift ($z = 3.572$) of the absorbing cloud used for measuring the D/H ratio. The quasar continuum at the $z = 3.572$ Lyman limit is therefore in transition, with the fluctuating absorption from numerous blended lines of individual Lyman-forest clouds becoming increasingly overlaid by the overlapping Lyman continua of the forest clouds. We have considered two ways of estimating the continuum level in the $z = 3.572$ Lyman continuum region. The first was a simple linear extrapolation of the estimated continuum at longer wavelengths which is shown as the dotted line in Fig. 2a. However, as the figure shows, even this very conservative procedure rules out the value of the neutral hydrogen column density ($N(\text{H I})$) reported in ref. 6. In the spectral interval $4,125 \text{ \AA} < \lambda < 4,175 \text{ \AA}$ the signal-to-noise weighted residual flux is 0.015 ± 0.014 D. Tytler, personal communication). This is not in disagreement with our spectra, as Fig. 2

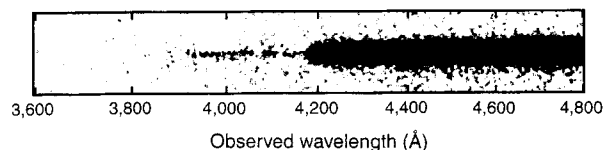


FIG. 1 The two-dimensional spectrum of Q1937 – 1009, shown in a grey-scale representation with the dispersion direction along the horizontal (covering the wavelength range 3,600–4,800 Å) and the spatial direction along the vertical (± 8.5 arcsec about the quasar). The Lyman continuum break is positioned at the centre of the figure with the residual flux at shorter wavelengths running to the left. The spectrum was formed from three 15-min exposures obtained with the 300 lines per mm grating on the LRIS spectrograph on the KeckI 10-m telescope using a 1.5×172 arcsec slit. The resulting spectrum had $R = 300$. The quasar was moved to a different position along the slit between each exposure. The frames were sky-subtracted using the median of positions from 8.5 to 4.5 arcsec from the quasar on either side of the quasar. The three frames were then median added to eliminate cosmic rays, and finally flux calibrated using observations of the white dwarf photometric standard star Feige 110, obtained in the same configuration. The spectrum was wavelength-calibrated using a third-order polynomial fit to observations of a krypton-mercury lamp, and the calibration checked against the Balmer line positions in the white dwarf.

shows that this region suffers particularly heavy absorption from the Lyman- α forest lines. However, the total column density adopted in ref. 6 ($N(\text{H I}) = 8.7 \times 10^{17} \text{ cm}^{-2}$) would predict that the residual flux below the Lyman break was $0.004^{+0.003}_{-0.002}$ which is shown as the dashed line and error bar, and lies well below the observed spectrum. In the wavelength region 890–900 Å we measure an average flux of 5.4 ± 0.4 counts compared to an extrapolated average continuum level of 224 counts, where the 1σ error is made by extracting sky regions in the same way as the quasar. This corresponds to an optical depth of $\tau = 3.72 \pm 0.06$ or $N(\text{H I}) = (5.9 \pm 0.1) \times 10^{17} \text{ cm}^{-2}$.

However, the statistical errors are small compared with the systematic error in the continuum placement, and the simple procedure used above overestimates the value of the optical depth and hydrogen column density. To provide an improved estimate, we have modelled the expected continuum level using statistical Ly α forest data^{12,13}. Such models provide a good representation of the average forest absorption, and we see that they also match the Q1937–1009 forest absorption rather well (dash-dot line in Fig. 2a). So long as there are no unrecognized strong absorption systems with redshifts slightly lower than $z = 3.572$, and the model gives an acceptable representation of the Ly α forest region, the Lyman continuum absorption required by the line absorption should predict the additional Lyman continuum absorption decrement to be applied to the quasar spectrum. In high-redshift Lyman- α forests, such as the one in Q1937–1009, the weak hydrogen clouds are so numerous that they blend into a quasi-smooth continuum. But Madau *et al.*¹³ noted that the contribution to the Ly- α optical depth, $\tau_{\text{Ly}\alpha}$, by optically thin absorbers with $10^{16} \leq N(\text{H I}) \leq 10^{17} \text{ cm}^{-2}$ is highly uncertain. In our calculation of $\tau_{\text{Ly}\alpha}$, we retained only the first term in their equation (5) and reduced its coefficient by 20%. This procedure assumes that there are no high column density hydrogen clouds in Q1937–1009 with redshifts near that of the $z = 3.572$ system. This assumption is justified both by the good match of the statistical model to the Ly α forest region of our spectra and by the finding of Tytler and Burles¹⁴ that the column densities of the three highest column density hydrogen clouds with redshifts to the blue of the D/H absorption system have $15.06 \leq \log N(\text{H I}) \leq 15.7$. As the combined absorption produced by these three systems is $\leq 10\%$ of Tytler's predicted absorption for the much stronger $z = 3.572$ D/H system it can be neglected in comparison to the uncertainties inherent in setting the appropriate level for the unabsorbed quasar continuum, and the accurate determination, with only low-resolution spectra, of the $z = 3.572$ Lyman continuum absorption in the presence of strong line absorption from other redshift systems. This fit reduces the extrapolated continuum level at 890–900 Å to 126 counts, which gives an optical depth of 3.15 ± 0.06 or $N(\text{H I}) = (5.0 \pm 0.1) \times 10^{17} \text{ cm}^{-2}$.

Figure 2b shows a linear plot of the moderate-resolution spectrum in the region near the $z = 3.572$ Lyman limit. Again the two continuum estimates are shown. Here Ly α forest lines can be seen cutting into the residual flux both above and below the break. The residual flux levels from the moderate- and low-resolution spectra are in satisfactory agreement when the moderate-resolution data is smoothed to the low resolution. Here we obtain $\tau_{\text{cont}} = 3.1$ (linear extrapolation of the quasar continuum) or $\tau_{\text{cont}} = 2.4$ (Ly α forest model continuum) where we have made the measurement at the rest-frame wavelength of 890 Å in the high resolution spectrum. These optical depths translate to $3.8 \times 10^{17} \text{ cm}^{-2} < N(\text{H I}) < 4.9 \times 10^{17} \text{ cm}^{-2}$ for the total hydrogen column density in the $z = 3.572$ absorption system.

Based on the above analysis of the two spectra, we consider a value of $5 \times 10^{17} \text{ cm}^{-2}$ to be a reasonable maximum neutral hydrogen column density that can be associated with the measured⁶ deuterium column density of $N(\text{D I}) = 2 \times 10^{13} \text{ cm}^{-2}$. This then gives a formal minimum for D/H of 4×10^{-5} .

However, this new determination of total H I column density cannot distinguish different H I or D I velocity components and so cannot measure the D/H ratio in any one component. Spectra with

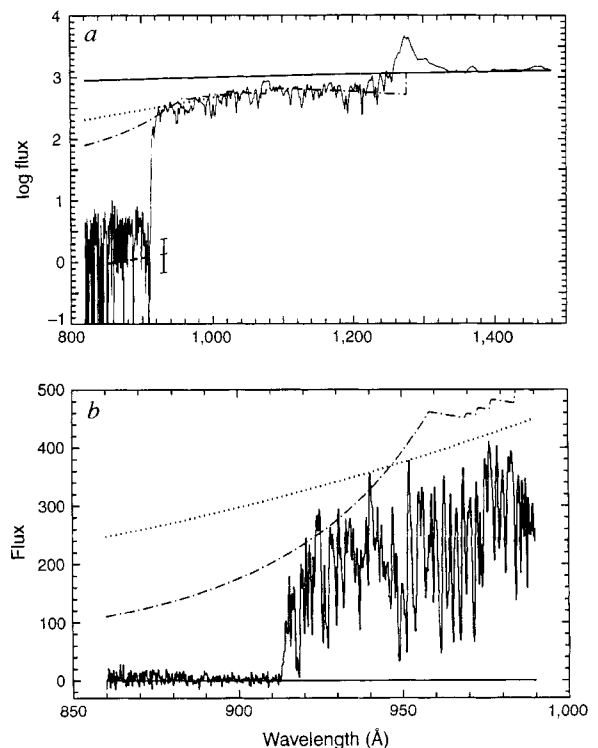


FIG. 2 a, The spectrum (flux per unit frequency in arbitrary units) of Q1937–1009 extracted from the two-dimensional spectral image of Fig. 1, shown as a function of the rest wavelength in the $z = 3.572$ quasar absorption line frame, showing the position of the Lyman continuum break. The solid line shows a fit to the upper envelope of the spectrum redward of the Ly α emission line, excluding known emission lines. The dash-dot line shows the level of the corresponding continuum expected blueward of the Ly α emission line from ref. 13. The dotted line shows a simple linear fit to the upper envelope of the continuum just above the Lyman continuum break. The dashed line shows the expected residual flux from the prediction of ref. 6. The error bar shows the uncertainty in that prediction, including both the quoted statistical and systematic errors. Both this spectrum and the higher-resolution spectrum (b) were oversampled and are plotted without smoothing. b, The higher-resolution spectrum ($R \approx 1500$), shown as flux per unit frequency (in arbitrary units) as a function of rest wavelength in the quasar absorption line frame. The spectrum was extracted in the same way as was the lower-resolution spectrum, and its spectrophotometry found to be in good agreement with the lower-resolution wide-slit spectrum. This higher resolution shows the rich forest of Ly α lines. The linear (dotted) and ref. 13 (dash-dot) continuum fits are again shown. The optical depth at 890 Å in the rest frame, where the signal-to-noise ratio is still high in the residual spectrum, was $\tau = 2.4$ for the continuum of ref. 13 and $\tau = 3.1$ for the linear fit.

a resolution $R \geq 6 \times 10^4$ are needed to resolve the velocity structure of the low-ionization metal line clouds in order to settle the issue of the number and relative strengths of the component clouds within the $z = 3.572$ system. More precise cloud information might show that a considerable fraction of the accompanying hydrogen could be located in the red wing of the line. If so, the corresponding deuterium would be obscured by overlying hydrogen, and D/H ratios even as high as 2×10^{-4} could then be accommodated by the data. We emphasize that the highly uncertain and model-dependent nature of the systematic errors involved in these measurements of deuterium makes the assignment of formal systematic errors very difficult. As this is likely to remain the case as new measurements accumulate, we have chosen rather to emphasise a reliable lower bound, while giving some idea through modelling of the possible uncertainty.

We therefore recommend adopting a range of $4 \times 10^{-5} <$

$D/H < 2.4 \times 10^{-4}$ as the most conservative current interpretation of the results of determining D/H in the high redshift quasar absorption lines systems. The density parameter, Ω_B , is then constrained to be $0.005 < \Omega_B h^2 < 0.016$, where the Hubble constant $H_0 = 100 h \text{ km s}^{-1} \text{ Mpc}^{-1}$. Unlike the lower D/H value of 2.3×10^{-5} , this range is broadly consistent with measurements of ^7Li and ^4He (ref. 8) and eliminates much of the evidence that suggests⁹ that there might be a crisis in the standard Big Bang nucleosynthesis model. It remains true that the upper end of the range is in better agreement with ^4He and ^7Li measurements whereas the lower value is easier to reconcile with models of Galactic chemical evolution tied to the locally measured value of D/H , as well as with local estimates of Ω_B . This issue will be resolved only when further measurements of D/H narrow the range. $\square$

Received 3 September; accepted 5 November 1996.

1. Songaila, A., Cowie L. L., Hogan, C. J. & Rugers, M. *Nature* **368**, 599–603 (1994).
2. Carswell, R. F. et al. *Mon. Not. R. Astron. Soc.* **268**, L1–L4 (1994).
3. Wampler, E. J. et al. *Astron. Astrophys.* **316**, 33–42 (1996).
4. Rugers, M. & Hogan, C. J. *Astrophys. J.* **459**, L1–L4 (1996).
5. Rugers, M. & Hogan, C. J. *Astron. J.* **111**, 2135–2140 (1996).
6. Tytler, D., Fan, X.-M. & Burles, S. *Nature* **381**, 207–209 (1996).
7. Burles, S. & Tytler, D. *Science* (submitted).
8. Fields, B. D., Kainulainen, K., Olive, K. A. & Thomas, D. *New Astron.* **1**(1), 77–96 (1996).
9. Steigman, G. *The Crisis Confronting Standard Big Bang Nucleosynthesis in Critical Dialogs in Cosmology* (Princeton Univ. Press, in the press).
10. Hata, N., Steigman, G., Bludman, S. & Langacker, P. *Phys. Rev. D*, (in the press).
11. Wampler, E. J. *Nature* **383**, 308 (1996).
12. Press, W. H. & Rybicki, G. *Astrophys. J.* **414**, 64–81 (1993).
13. Madau, P. et al. Preprint STSc-1085 (Space Telescope Science Institute).
14. Tytler, D. & Burles, S. in *Origin of Matter and Evolution of Galaxies in the Universe '96* (eds Kajino, T., Yoshii, Y. & Kubono, S.) (World Scientific, Singapore, in the press).

ACKNOWLEDGEMENTS. We thank E. Hu for her assistance in obtaining the data. The authors were visiting astronomers at the W. M. Keck Observatory, jointly operated by the California Institute of Technology and the University of California. This work was supported at the University of Hawaii by the State of Hawaii.

CORRESPONDENCE should be addressed to A.S. (e. mail: acowie@uhifa.ifa.hawaii.edu).

A 120-Mpc periodicity in the three-dimensional distribution of galaxy superclusters

J. Einasto*, M. Einasto*, S. Gottlöber†, V. Müller‡, V. Saar*, A. A. Starobinsky‡, E. Tago*, D. Tucker§, H. Andernach|| & P. Frisch¶

* Tartu Observatory, EE2444 Tõravere, Estonia

† Astrophysical Institute Potsdam, An der Sternwarte 16, D-14482 Potsdam, Germany

‡ Landau Institute Theoretical Physics, 117946 Moscow, Russia

§ Fermilab, Batavia, Illinois 60510, USA

|| INSA, ESA JUE Observatory, E-28080 Madrid, Spain, and Depto. de Astronomía, Univ. Guanajuato, Guanajuato, Mexico

¶ University Observatory, D 37083 Göttingen, Germany

ACCORDING to the favoured models for the formation of large-scale structure in the Universe (in which the dynamics of the Universe is dominated by cold dark matter), the distribution of galaxies and clusters of galaxies should be random on large scales. It therefore came as a surprise when a periodicity was reported¹ in the distribution of high-density regions of galaxies in the directions of the Galactic poles, although the apparent lack of periodicity in other directions led to the initial report being regarded as a statistical anomaly². A subsequent study^{3–6} also claimed evidence for periodicity on the same scale, but the

statistical significance of this result was uncertain due to the small number of clusters used. Here, using a new compilation⁷ of available data on galaxy clusters, we present evidence for a quasi-regular three-dimensional network of rich superclusters and voids, with the regions of high density separated by $\sim 120 \text{ Mpc}$. If this reflects the distribution of all matter (luminous and dark), then there must exist some hitherto unknown process that produces regular structure on large scales.

During the past few years the number of clusters with measured redshifts has increased considerably. To search for the possible presence of a regularity of the distribution of matter in the Universe we have used a new compilation⁷ of available data on rich clusters of galaxies catalogued by Abell and collaborators^{8,9}. The compilation has made use of all (~ 300) published references on redshifts of both individual galaxies and Abell galaxy clusters. Individual galaxies were associated with a given Abell cluster if they lay within a projected distance of $\leq 1.5 h^{-1} \text{ Mpc}$ (1 Abell radius) and within a factor of two of the redshift estimated from the brightness of the cluster's tenth-brightest galaxy, using the photometric estimate of Peacock and West¹⁰ (h is the Hubble constant in units of $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$). The compilation contains measured redshifts for 869 of the 1304 clusters with an estimated redshift up to $z = 0.12$. For the present analysis we used all rich clusters (richness class $R \geq 0$) in this compilation with at least two galaxy redshifts measured. The omission of the 435 clusters without measured redshifts does not affect our result because an appropriate selection function was used.

This cluster sample (including clusters with estimated redshifts) was used to construct a new catalogue of 220 superclusters of galaxies^{11,12}. These are systems of clusters where the distances between nearest-neighbours among member clusters do not exceed $24 h^{-1} \text{ Mpc}$. In this way we find high-density regions in the distribution of clusters of galaxies. In Fig. 1 we plot clusters located in rich superclusters with at least eight member clusters. We see a moderately regular network of superclusters and voids with a step size of $\sim 120 \pm 20 h^{-1} \text{ Mpc}$ where chains of superclusters are separated by voids of almost equal size^{11,12}. The whole distribution resembles a three-dimensional chessboard¹³. The nearest-neighbour test, and pencil-beam and void analysis, indicate that clusters in poor superclusters with less than eight members and isolated clusters form a more uniform population, preferentially located in void walls between rich superclusters but not filling the voids¹².

To quantify the regularity of the cluster distribution we have calculated the correlation function and the power spectrum for clusters of galaxies. The correlation function describes the distribution of clusters in real space, the power spectrum in the Fourier space of density waves. Analysis of various geometric models has shown that if superclusters form a quasi-regular lattice with an almost constant step size then the cluster correlation function is oscillating, and it has alternate secondary maxima and minima, separated by half the period of oscillations. The period of spatial oscillations of the correlation function is equal to the step size of the distribution¹⁴. The amplitude of the power spectrum at the wavelength corresponding to that period is enhanced with respect to other wavelengths, that is, it is peaked. In contrast, if superclusters are located randomly in space then the correlation function approaches zero level at large separations, and the power spectrum turns smoothly from the region with positive spectral index on large wavelengths to a negative index on small wavelengths¹⁴.

To calculate the power spectrum we have used the sample which contains clusters with measured redshifts only and lying in both Galactic hemispheres out to the distance covered by our cluster and supercluster catalogues. The power spectrum was derived using two different methods: a direct method where we calculate the distribution of clusters in the wavenumber space, and an indirect method where we first calculate the correlation function of clusters of galaxies and then find the spectrum. In the latter case we make use of the fact that the power spectrum and the

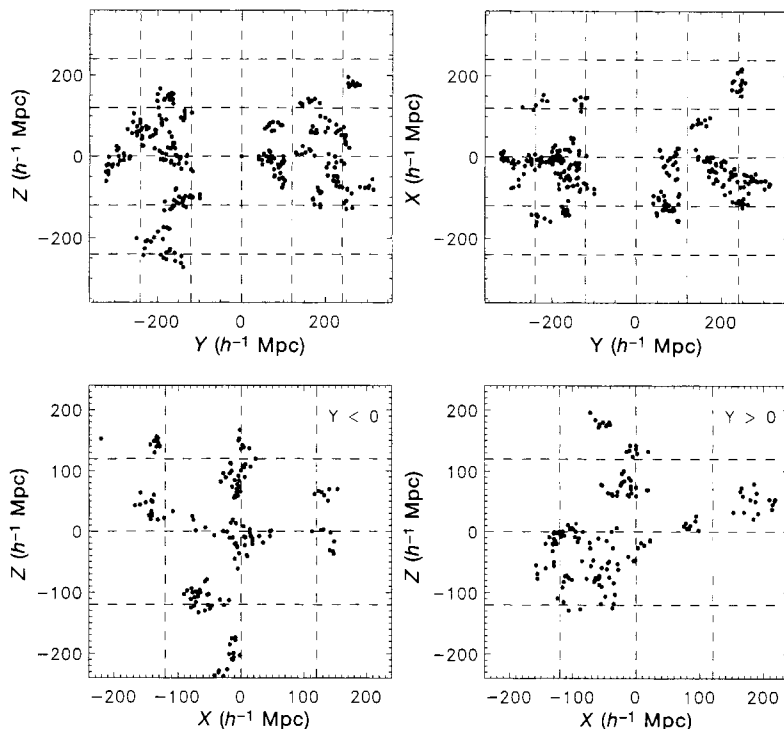


FIG. 1 The distribution of 319 clusters in 25 very rich superclusters with at least eight members (including 58 clusters with photometric distance estimates), illustrating the network in the cluster distribution in supergalactic coordinates. In the lower panels, clusters in the northern and southern Galactic hemispheres are plotted separately. The supergalactic $Y = 0$ plane coincides almost exactly with the Galactic equatorial plane, that is, with the zone of avoidance due to galactic absorption. The grid with step size $120 h^{-1} \text{ Mpc}$ corresponds approximately to distances between high-density regions across voids. In the two upper panels and in the lower right panel, several superclusters overlap due to project but are actually well-separated in space.

correlation function are related by the Fourier transform.

In both methods the main problem is the calculation of the selection function of clusters of galaxies which corrects for incompleteness both at low Galactic latitude b and at large distances r from the observer. The selection function can be represented by linear functions of $\sin b$ and r (for numerical values of selection function parameters see refs 12 and 15). Both methods of deriving the power spectrum yield similar results. However, parameters of the correlation function are insensitive to small inaccuracies of the selection function¹⁵, thus the indirect method yields more accurate results for the power spectrum.

The power spectrum for clusters of galaxies is shown in Fig. 2. This power spectrum is among the first for three-dimensional data having measurements on scales well above $100 h^{-1} \text{ Mpc}$. On very large scales the errors are large due to incomplete data. On moderate scales we see one single well defined peak at a wavenumber $k_0 = 0.052 h \text{ Mpc}^{-1}$. Errors are small near the peak, and the relative amplitude and position of the peak are determined quite accurately. The wavelength of the peak is $\lambda_0 = 2\pi/k_0 = 120 \pm 15 h^{-1} \text{ Mpc}$. Near the peak, there is an excess in the amplitude of the observed power spectrum over that of the cold dark matter (CDM) model (see below) by a factor of 1.4. Within observational errors our power spectrum on large scales is compatible with the Harrison-Zeldovich spectrum with constant power index $n = 1$, and on small scales with a spectrum of constant negative power index $n = -1.8$.

The power spectrum is often used to compare the distribution of matter in the Universe with theoretical predictions. Currently popular structure formation theories based on the dynamics of a Universe dominated by CDM yield spectra of which one example is plotted in Fig. 2 as a solid line. The spectrum is rising on long wavelengths λ (small values of the wavenumber $k = 2\pi/\lambda$), and falling at short wavelengths (large values of k). The transition between short- and long-wavelength regions in the CDM-spectrum is smooth. The distribution of superclusters in CDM-models is irregular¹⁶.

As we see, the relative amplitude of the observed power spectrum above the standard CDM-type model is not very large. Thus we may ask the question: within the framework of the standard cosmogony, how frequently can we expect to find a distribution of clusters which has a power spectrum similar to

that observed? To answer this question we determined the correlation function and power spectrum for clusters in rich superclusters of CDM-type models. In the spectral range of interest the power spectrum of the standard CDM model is similar to the spectrum of a random supercluster model¹⁴. We make use of this similarity by generating 1,000 realizations of the random supercluster model, applying the selection effects as found in

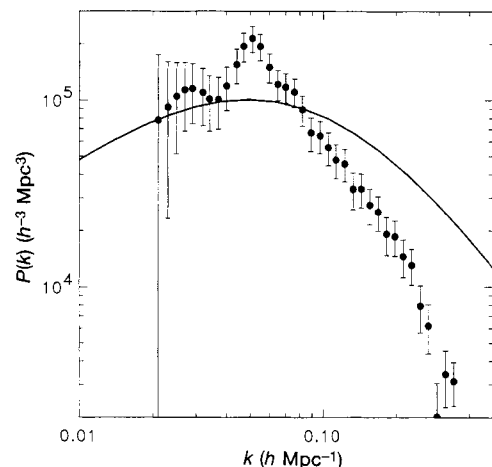


FIG. 2 The power spectrum $P(k)$ of 869 clusters with measured redshifts (filled circles). The spectrum was calculated from the cluster correlation function via the Fourier transform; both functions are found in redshift space. Errors were determined from 2σ errors of the correlation function found from the scatter of different simulations. For each wavenumber, we integrated over the whole interval of the correlation function so individual values of the spectrum are not independent of each other. To check the indirect method of calculation of the power spectrum we have used simulated cluster samples having similar selection effects as real ones. This check was performed for a wide variety of models with different initial spectra. Our results show that the true spectrum can be restored over the wavenumber interval from $k \approx 0.03 h \text{ Mpc}^{-1}$ towards shorter waves until $k \approx 0.3 h \text{ Mpc}^{-1}$. The solid line is the standard CDM ($h = 0.5$, $\Omega = 1$) power spectrum enhanced by a bias factor of $b = 3$ over the four-year COBE normalization.

cluster distribution, and determining the parameters of the cluster correlation function and power spectrum. We measure the mean period and amplitude of the oscillating correlation function and their respective scatter. We also calculate the deviations for individual periods. This test shows that combination of parameters (for instance, period versus amplitude) close to the observed values occurs in $\sim 1\%$ of cases, but the simultaneous concurrence of all parameters with observations is a much more rare event. Thus some change in the initial spectrum of matter is necessary in order to explain the observed correlation function and power spectrum for clusters of galaxies.

The regularity of the distribution of superclusters is quite striking and has some similarity with the regularity found in pencil-beam galaxy surveys¹. Independent evidence for the presence of a preferred scale in the Universe around $100h^{-1}$ Mpc comes from a core-sampling analysis of the distribution of sheet-like and filamentary structures¹⁷, and from a power spectrum analysis of the two-dimensional distribution of galaxies¹⁸ in the Las Campanas redshift survey. Available data, however, are insufficient to say whether one-, two-, and our three-dimensional surveys measure identical physical scales in the Universe, or, if so, whether differences in numerical values of the scale are due to differences in methods applied in determination or to a real cosmic variance in the scale.

According to our present understanding of the Universe, the observed distribution of luminous matter is strongly correlated with the density perturbations of some dark non-baryonic matter component(s) at the moment of recombination. These perturbations are due to processes in the very early Universe so that their power spectrum at recombination depends both on these pro-

cesses and the further evolution of the Universe, in particular the transition from radiation- to matter-dominated expansion. We have shown that the luminous matter in galaxy clusters is much more regularly distributed than expected. Thus we conclude that our present understanding of structure formation on very large scales needs revision. $\square$

Received 10 October; accepted 19 November 1996.

1. Broadhurst, T. J., Ellis, R. S., Koo, D. C. & Szalay, A. S. *Nature* **343**, 726–728 (1990).
2. Kaiser, N. & Peacock, J. A. *Astrophys. J.* **379**, 482–506 (1991).
3. Kopylov, A. I., Kuznetsov, D. Y., Fetisova, T. S. & Shvarzman, V. F. in *Large Scale structure of the Universe* (eds Audouze, J., Pelletan, M.-C. & Szalay, A.) 129–137 (Kluwer, Dordrecht, 1988).
4. Mo, H. J., Deng, Z. G., Xia, X. Y., Schiller, P. & Börner, G. *Astron. Astrophys.* **257**, 1–10 (1992).
5. Fetisova, T. S., Kuznetsov, D. Y., Lipovetsky, V. A., Starobinsky, A. A. & Olowin, R. P. *Astron. Lett.* **19**, 198–202 (1993).
6. Einasto, J. & Gramann, M. *Astrophys. J.* **407**, 443–447 (1993).
7. Andemach, H., Tago, E. & Stengler-Larrea, E. *Astrophys. Lett. Commun.* **31**, 27–30 (1995).
8. Abell, G. *Astrophys. J. Suppl. Ser.* **3**, 211–288 (1958).
9. Abell, G., Corwin, H. & Olowin, R. *Astrophys. J. Suppl. Ser.* **70**, 1–138 (1989).
10. Peacock, J. & West, M. J. *Mon. Not. R. Astron. Soc.* **259**, 494–504 (1992).
11. Einasto, M., Einasto, J., Tago, E., Dalton, G. & Andemach, H. *Mon. Not. R. Astron. Soc.* **269**, 301–322 (1994).
12. Einasto, M., Tago, E., Jaaniste, J., Einasto, J. & Andemach, H. *Astron. Astrophys.* (in the press); SISSA preprint astro-ph 9610088 (1996).
13. Tully, R. B., Scaramella, R., Vettolani, G. & Zamorani, G. *Astrophys. J.* **388**, 9–16 (1992).
14. Einasto, J. et al. *Mon. Not. R. Astron. Soc.* (submitted).
15. Einasto, J. et al. *Mon. Not. R. Astron. Soc.* (submitted).
16. Frisch, P. et al. *Astron. Astrophys.* **296**, 611–627 (1995).
17. Doroshkevich, A. G. et al. *Mon. Not. R. Astron. Soc.* (submitted).
18. Landy, S. D. et al. *Astrophys. J.* **456**, L1–L4 (1996).

ACKNOWLEDGEMENTS. We thank P. J. E. Peebles for discussion and valuable suggestions. This work was supported by the International Science Foundation and the Estonian Science Foundation; visits of J. E. and A. S. to Potsdam were supported by the Deutsche Forschungsgemeinschaft. A.S. was partially supported by the Russian Research Project 'Cosmomicrophysics'.

CORRESPONDENCE should be addressed to J.E.

Entropy difference between the face-centred cubic and hexagonal close-packed crystal structures

L. V. Woodcock

Department of Chemical Engineering, University of Bradford, Bradford BD7 1DP, UK

SPHERES can be stacked into two close-packed crystalline arrangements, face-centred cubic (f.c.c.) and hexagonal close-packed (h.c.p.), which have identical close-packed volumes¹ and very similar equations of state². But they are structurally distinct, implying that they might have different thermodynamic properties and stabilities. Finding a difference in free energy between the two structures has been the objective of much theoretical^{3,4} and computational^{5–7} effort, but without a conclusive resolution. Here I report that a significant difference in the pressure–volume (P – V) behaviour can be detected in the vicinity of a mechanical instability point within a single-occupancy cell model⁸ of these packings. This model provides an exact thermodynamically reversible path between the two structures, and so the P – V isotherms can be integrated to obtain the Gibbs free-energy difference. I find that the f.c.c. phase is more stable by around $0.005RT$ per mol (where R is the universal gas constant); as the enthalpy difference is negligible, this implies that the entropy difference is of the order of $0.005R$ for all temperatures up to the melting point.

Both these close-packed crystal structures occur extensively in nature; 25% of the elements crystallize in the f.c.c. structure and 20% of the elements crystallize h.c.p. Thermodynamic equi-

lium requires that crystallizing atomic or molecular fluids adopt the structure with the lowest Gibbs free energy under external conditions of constant temperature (T) and pressure (P). It has been known for many decades that a classical fluid of hard spheres in thermodynamic equilibrium will crystallize under sufficient pressure; the crystal structure of the stable solid phase, however, has remained unknown.

To establish a fluid–crystal transition pressure in the hard-sphere model, Hoover and Ree⁹ calculated the communal entropy of the face-centred cubic single-occupancy (SO)-cell model, as first defined by Kirkwood⁸. In this model each sphere centre is confined within the walls of its primary Voronoi polyhedron (that is, the local cell around a lattice site, bounded by the planes that bisect the lines joining the site to its neighbouring sites). For both the f.c.c. and the h.c.p. crystal structures all the primary cells are regular dodecahedral in shape, each of the 12 faces representing the perpendicular plane bisecting the line between first-neighbour sites. For N spheres there are N space-filling dodecahedra, each containing a single sphere, and all of which are identical. This model provides a reversible path from crystal to fluid which enables the relative free energy to be calculated and hence the crystallization transition pressure to be located. For pressures above freezing, the unconstrained crystal is stable without the constraining walls and the thermodynamic properties are indistinguishable from the SO-cell model. At very low, ideal-gas, pressures the difference in molar entropy between the free fluid phase and the SO-cell model is exactly equal to R .

Kirkwood's communal entropy ($\Delta_{\text{comm}}S_{T,V}$) at higher density is the entropy difference between the single-occupancy cell model and the free fluid at the same temperature and volume. Physically, communal entropy represents the additional entropy a model fluid acquires when its molecules are free to explore the whole available volume, compared to a crystal or glassy state in which the molecules are naturally constrained. One could equally define an alternative communal entropy as the entropy difference between the SO-cell model and the free fluid at the same temperature and

pressure ($\Delta_{\text{com}}S_{T,P}$). The two communal entropies are exactly related by

$$\Delta_{\text{com}}S_{T,V} = \Delta_{\text{com}}S_{T,P} + R \ln Z \quad (1)$$

where $Z = PV/RT$. In the low-density limit, $Z \rightarrow 1$, $\ln Z \rightarrow 0$, and therefore $\Delta_{\text{com}}S_{T,V} = \Delta_{\text{com}}S_{T,P} = R$.

In the hard-sphere model the communal Gibbs free energy is

$$G_{\text{fluid}} - G_{\text{SO}} = \Delta_{\text{com}}G_{T,P} = P\Delta V_{T,P} - T\Delta_{\text{com}}S_{T,P} \quad (2)$$

where G_{fluid} and G_{SO} are the Gibbs free energies of fluid and SO-cell model, respectively and $V_{T,P}$ can be regarded as a 'communal volume', that is, the excess volume of the free fluid over the SO-cell model at the same temperature and pressure. It approaches the constant value of $2\pi N\sigma^3/3$ (where σ is the sphere diameter) in the ideal gas limit, as $P \rightarrow 0$. At constant T , and vanishing P , therefore, the Gibbs free-energy difference between the SO-cell model and the free fluid is exactly RT .

For any reversible process, from the first and second laws of thermodynamics, Gibbs free energy changes with temperature and pressure according to

$$dG = VdP - SdT \quad (3)$$

where V is the volume and S is the entropy. The pressure at which the hard-sphere fluid crystallizes (P_i) can be simply obtained by integrating equation (3). Along any P - V isotherm the SdT term drops out and the freezing pressure is obtained from the condition

$$\Delta G_{\text{fluid-SO}} = \int_0^{P_i} \Delta V dP - RT = 0 \quad (4)$$

Given the equations of state of both the fluid from freezing to the ideal gas (which is well represented by the Carnahan-Starling equation¹⁰) and the SO-cell model from low density to crystal close packing, the freezing pressure of the hard-sphere fluid is determined (Fig. 1a).

The SO-cell model undergoes a phase transition at the pressure P_i^* where the walls would become necessary to maintain even transient crystal-like stability of the otherwise free crystal. Up to this pressure, the equation of state is given by

$$P_{\text{SO}}^* = \left[1 + \pi\sqrt{2}\left(\frac{V_0}{V}\right)^{\frac{4}{3}} + \left(\pi\sqrt{2}\right)^2 \left(\left(\frac{V_0}{V}\right)^3 + \left(\frac{V_0}{V}\right)^6 \right) \right] / V^* \quad (5)$$

where P_{SO}^* is the reduced pressure in hard-sphere units of $k_B T/\sigma^3$ (k_B is Boltzmann's constant), V^* is the reduced volume in units of $N\sigma^3$ and V_0 is the minimum volume at close-packing. Beyond the first two terms, equation (5) is empirical, and determined from the present molecular-dynamics (MD) SO-cell data; it is accurate up to the structural instability transition pressure ($P_i^* = 8.5$) in the SO-cell model.

At pressure above this structural transition point (Fig. 1a) the equation of state of both the SO-cell model and the free crystal are accurately represented, from P_i^* up to infinite pressure (close packing), by the expansion^{2,11}

$$V_{\text{CRYST}}^* = \frac{1}{\sqrt{2}} + 3P^{*-1} - \frac{4}{3}\sqrt{2}P^{*-2} + 4P^{*-3} + 36\sqrt{2}P^{*-2} \quad (6)$$

The pressure of the freezing transition was then calculated analytically, by integrating the communal volume, and found to be 12.06. The coexistence packing fractions at melting and freezing are 54.3% (from equation (6)) and 48.9% (from the Carnahan-Starling equation¹⁰) respectively. This new determination is within the uncertainty of Hoover and Ree's Monte Carlo result⁹.

The SO-cell model also provides a reversible thermodynamic path between alternative crystal phases, f.c.c. and h.c.p., because, in this model, the low-density thermodynamic properties of the two structures are identical. The free energy difference between the f.c.c. and h.c.p. crystalline phases at crystallization may be computed by integration of the respective equations of state, if any significant pressure difference could be accurately established. The question is, where is the largest pressure difference between the f.c.c. and h.c.p. SO-cell models most likely to be found?

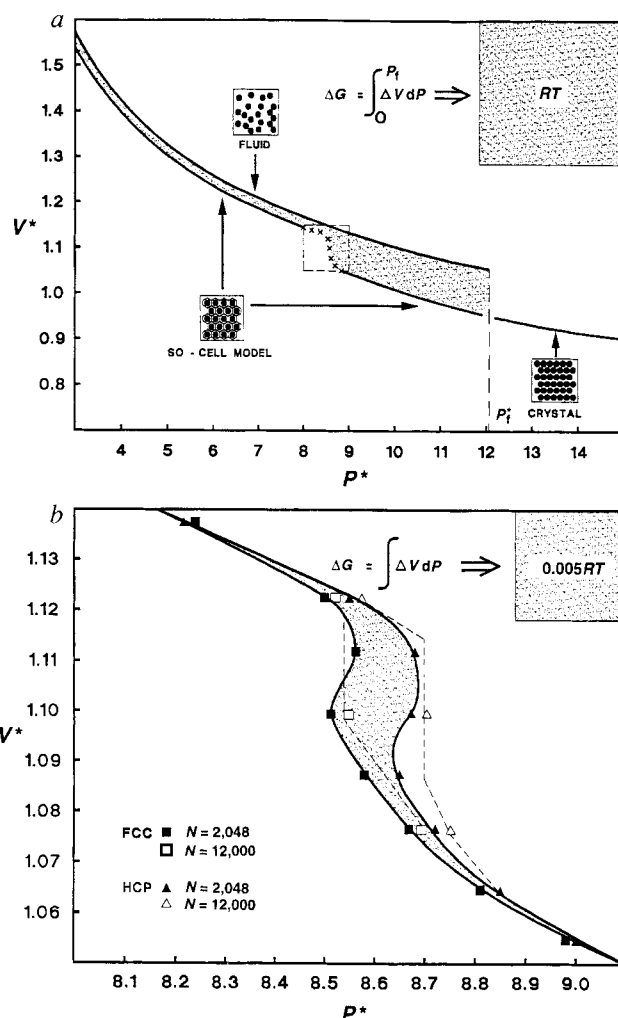


FIG. 1 Calculation of the freezing crystallization parameters of hard spheres. a, Graphical representation of the equations of state of the hard-sphere fluid¹⁰, the single-occupancy (SO)-cell model (equation (5)) and the unconstrained crystal phase (equation (6)). (Here P^* is the reduced pressure, V^* is the reduced volume—see text.) The freezing pressure (P_i) is calculated by matching the shaded area of integration, which represents a free energy, with the communal free-energy difference at zero pressure (RT) represented as the shaded square in the top right-hand corner. The SO-cell model shows a mechanical instability transition around $P^* \approx 8.5$; it is in this region that the origin of the f.c.c.–h.c.p. free-energy difference is to be found. The state points chosen for extensive MD computation, for the single occupancy model, are illustrated by crosses. A very close inspection of more accurate data (b) for h.c.p. and f.c.c. structures reveals a substantial difference in pressures and hence crystal-phase free energies. b, Data for the volume–pressure behaviour of the f.c.c. and h.c.p. SO-cell models from MD computations of hard spheres in the vicinity of the instability transition; the results show that, over this narrow range, the f.c.c. pressure is significantly less than the h.c.p. pressure, and that at the actual point of the transition, the two structures have different instability pressures. A numerical integration of the volume difference with respect to pressure, represented by the shaded area between the respective equation of state data, yields a free energy area equal to $0.005RT$ as illustrated by the equivalent area in the top right-hand corner. The vertical dashed lines suggest the transition behaviour for the thermodynamic limit of large systems where $N = \infty$, from which the numerical result for the entropy difference is estimated. The box dimensions, in multiples of a cell dimension, used in the computations are as follows. (1) For $N = 2,048$ with four spheres per unit volume. f.c.c.; $x, 8d_0; y, 8d_0; z, 8d_0$; h.c.p.; $x, 8d_0; y, 8\sqrt{3}d_0; z, 8\sqrt{8/3}d_0$. (2) For $N = 12,000$ x, y layers with two spheres per unit volume, stacked in z -direction. f.c.c.; $x, 20d_0; y, 10\sqrt{3}d_0; z, (8\sqrt{8}d_0)/3$; h.c.p.; $x, 20d_0; y, 10\sqrt{3}d_0; z, 30\sqrt{2/3}d_0$. Note that for $N = 12,000$ the box dimensions are identical. Here d_0 is the lattice parameter.

The f.c.c. and h.c.p. lattices differ only in the third and successive neighbours. The coordination numbers (n) up to fifth neighbours are given in Table 1. For a thermodynamic hard-sphere fluid, the pressure at a fixed temperature and volume is directly related to the collision frequency by

$$P^* = \left(1 + \frac{n_c(\tau)\sqrt{\pi}}{3\tau}\right)/V^* \quad (7)$$

wherein n_c is the number of collisions in the time τ . In the stable crystal region only first neighbours collide; one would not expect to find a significant pressure difference in the stable crystal range up to the melting transition because of first-neighbour collision dominance. Likewise, in the lower density region of the SO-cell model, the two respective f.c.c. and h.c.p. structures have identical free energies relative to the ideal gas (RT) and exactly the same low-density equation of state given by the first two terms in equation (5).

Lengthy MD calculation, of the order of 10 million collisions, of the pressure in the SO-cell model in the intermediate density range, however, reveal that there is a region of mechanical instability in which there are significant contributions to pressure from second, third, and, in the h.c.p. case, fourth and fifth neighbour collisions, which differ between f.c.c. and h.c.p. This mechanical instability is a non-physical artefact of the SO-cell nuclei; it appears to be a first-order phase transition from a quasi-crystalline state in which the spheres are localized at the cell centres, to a quasi-fluid state in which the singlet probability density is delocalized within each cell. It is in this region of the mechanical instability transition that this small, but significant, difference in pressure between the two structures is to be found. Extensive MD computations in the SO-cell model, for both f.c.c. and h.c.p. structures, with systems of 500, 2,048 and 12,000 spheres, averaging between 10 million and 100 million collisions for each data point, have been carried out at in the volume range $V^* = 1.00$ – 1.25 .

Figure 1b shows a detailed plot of these pressure data (for the larger systems); it is a $100\times$ expansion of the small dashed square in Fig. 1a. This high-resolution picture of the respective equation-of-state shows clearly that the two structures have quite different properties in the vicinity of the SO-cell model structural instability transition. Moreover, there is a small pressure difference in the immediate region on both sides of the transition of approximately $0.02k_B T/\sigma^3$. On the low-pressure side of the transition, the h.c.p. has the lower pressure, whereas on the high-pressure side of the transition, the f.c.c. has the lower pressure. Interestingly, after numerical integration, these small tail contributions on each side of the transition largely cancel.

Both transitions in the respective SO-cell models show characteristics of a first-order phase transition; an inversion loop for the smaller systems, and transient metastability on either side of the transition. The data in Fig. 1b have been obtained mainly for a system of $N = 2,048$ spheres, with period boundary conditions.

Computations have also been carried out for $N = 500$ and $N = 12,000$ to confirm the interpretation of the present data. The transition pressures for infinite lattices are estimated to be 8.55 and 8.70 for the two structures respectively.

The Gibbs free-energy difference between the two phases is obtained by integrating the area between the f.c.c. and h.c.p. single occupancy isotherms. The result is $G_{\text{hcp}} - G_{\text{fcc}} = 0.005RT$; an equivalent shaded square area is shown in Fig. 1b. Frenkel and Ladd⁷ found no free-energy difference within the uncertainty of their computations ($0.0015RT$, which is less than the difference of $0.005RT$ obtained here). Both their method and mine devise a reversible path connecting the two structures to an idealized state, where the free energy is analytic and the same for both structures. Frenkel and Ladd, however, integrate the free-energy difference over the whole path, because their method cannot identify the physical origin of any such difference. In the single-occupancy model, on the other hand, the source of the difference is a very small region of the whole path, and by focusing the calculation on this region I have been able to quantify this difference.

These computations are being continued to improve the accuracy. From the present SO-cell equation-of-state data, however, it seems clear that the difference in thermodynamic properties between these two crystal structures is significant, amounting to $0.005R$ with an estimated uncertainty of $\sim 20\%$ for the molar entropy. The f.c.c. crystal structure of atomic hard spheres is more stable than the h.c.p. structure by a significant molar Gibbs free-energy difference of $\sim 0.005RT$. Further computations are underway to reduce the small uncertainties arising from the numerical integrations, to assess the effect of imposing isotropy on the h.c.p. structure which has only two-fold symmetry compared to the six-fold f.c.c., and further, to calculate the free energies of alternative hybrid structures of the stacking type-ABCAB-and so on which have not previously been considered at all. None of these considerations, however, are likely to alter the present conclusion that the stable structure of crystalline spheres is face-centred cubic. □

Received 22 July; accepted 11 November 1996.

1. Kittel, C. *Introduction to Solid State Physics* 5th Edn, Ch. 1 (Wiley, New York, 1976).
2. Alder, B. J., Hoover, W. G. & Young, D. A. *J. Chem. Phys.* **49**, 3688–3696 (1968).
3. Stillinger, F. H. & Salzburg, Z. W. *J. Chem. Phys.* **46**, 3962–3975 (1967).
4. Rudd, W. G., Salzburg, Z. W., Yu, A. P. & Stillinger, F. H. *J. Chem. Phys.* **49**, 4857–4863 (1968).
5. Alder, B. J., Carter, B. P. & Young, D. A. *Phys. Rev.* **183**, 831–833 (1969).
6. Alder, B. J., Young, D. A., Mansigh, M. R. & Salzburg, Z. W. *J. Comput. Phys.* **7**, 361–366 (1971).
7. Frenkel, D. & Ladd, A. J. C. *J. Chem. Phys.* **81**, 3188–3193 (1984).
8. Kirkwood, J. G. *J. Chem. Phys.* **18**, 380–382 (1950).
9. Hoover, W. G. & Ree, F. H. *J. Chem. Phys.* **49**, 3609–3619 (1968).
10. Carnahan, N. F. & Starling, K. E. *J. Chem. Phys.* **51**, 635–638 (1979).
11. Woodcock, L. V. *Ann. NY Acad. Sci.* **371**, 274–298 (1981).

CORRESPONDENCE should be addressed to L.V.W. (e-mail: l.v.woodcock@bradford.ac.uk).

TABLE 1 Properties of the first five coordination shells

		Neighbour				
		1st	2nd	3rd	4th	5th
f.c.c.	Number of neighbours	12	6	24	12	24
	Separation/ d_0	1	$2^{1/2}$	$3^{1/2}$	$4^{1/2}$	$5^{1/2}$
h.c.p.	Number of neighbours	12	6	2	18	12
	Separation/ d_0	1	$2^{1/2}$	$(8/3)^{1/2}$	$3^{1/2}$	$(11/3)^{1/2}$

Coordination numbers of the first five coordination shells in the face-centred cubic (f.c.c.) and hexagonal close-packed (h.c.p.) lattice structures. In the single-occupancy (SO)-cell model, at intermediate densities around the vicinity of the phase transition, collisions can occur between third neighbours in the f.c.c. structure, and up to fifth neighbours in the h.c.p. structure.

Single-molecule spectral fluctuations at room temperature

H. Peter Lu & X. Sunney Xie

Pacific Northwest National Laboratory, Environmental Molecular Sciences Laboratory, PO Box 999, Richland, Washington 99352, USA

RECENT advances in near-field¹ and far-field^{2,3} fluorescence microscopy have made it possible to image single molecules and measure their emission^{3,4} and excitation⁵ spectra and fluorescence lifetimes^{3,6–8} at room temperature. These studies have revealed spectral shifts⁴ and intensity fluctuations^{6,7}, the origins of which are not clear. Here we show that spontaneous fluctuations in the spectra of immobilized single dye molecules occur on

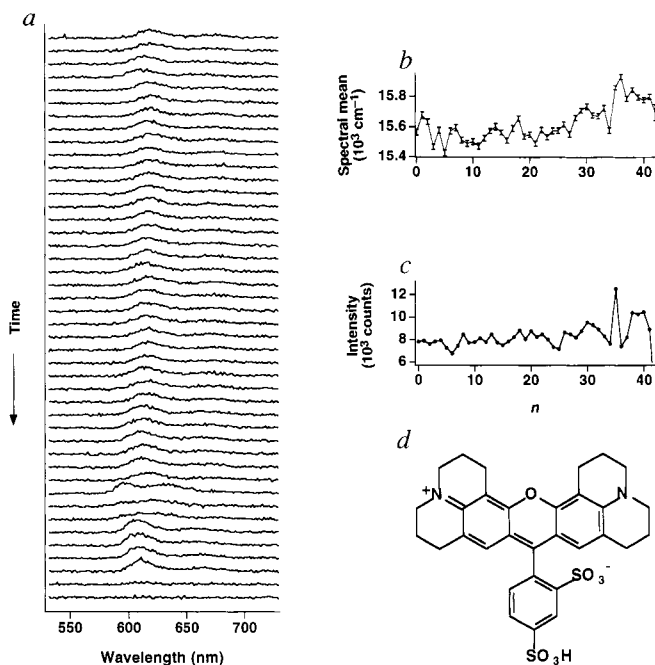


FIG. 1 a, Emission spectra of a single sulphorhodamine 101 molecule taken sequentially with 170-ms data collection times before the final photobleaching. Significant spectral shifts are evident. b, Trajectory of the spectral mean (in wavenumbers) of the molecule; n is the index number of the spectra. The fluctuations are clearly beyond the error bars. c, Trajectory of the total emission intensity (integrated area of each spectrum) of the molecule. d, Structure of the molecule. The emission spectra of individual molecules were recorded with a high emission collection efficiency (10%) using a combination of a spectrograph (Acton Research, Acton) and a back-illuminated CCD camera (Princeton Instruments, Trenton) configured for a spectral resolution of 1 nm with no dead time in data collection.

two different timescales: hundreds of milliseconds and tens of seconds, indicating that these fluctuations have two distinct activation energies. In addition, we see photoinduced spectral fluctuations on repeated photoexcitation of single molecules. We suggest that all of these fluctuations can be understood as transitions between metastable minima in the molecular potential-energy surface.

We obtained diffraction-limited fluorescence images of individual sulphorhodamine 101 molecules with a modified inverted fluorescence microscope (Nikon), in a similar fashion to a recent report³. The sample was prepared by first spin-coating a quartz surface with a 10^{-9} M dye solution and then with a 20-nm polymethylmethacrylate film to prevent fast photobleaching. Figure 1a shows the emission spectra of a single molecule sequentially recorded with 170-ms collection times. The trajectory of the spectral means is shown in Fig. 1b. The fluctuations of the spectral means are evident, while the spectral widths remain essentially constant. In addition, the total emission intensity (that is, the integrated area of each spectrum) fluctuates with time (Fig. 1c) and is consistent with other observations using near-field microscopy^{6,7}. The intensity fluctuations were previously attributed to spectral shifts.⁶ Indeed, the intensity fluctuations (Fig. 1c) essentially correlate with the spectral fluctuations in Fig. 1b: a blue-shifted emission (and absorption) spectrum results in a larger absorption cross-section at the excitation wavelength (532 nm, at the blue edge of the absorption band). The spectra shown in Fig. 1 were taken

with an excitation rate of $5 \times 10^5 \text{ s}^{-1}$, which is much lower than the saturation rate⁹. At saturation, intensity fluctuation arises from a different origin: trapping into the metastable triplet state^{10,11}.

Every molecule that we examined in a large population exhibited similar spectral fluctuations. To evaluate these statistically, we analysed the autocorrelation function $C(t) = \langle v(0)v(t) \rangle - \langle v \rangle^2$ of spectral trajectories, v being the spectral mean in wavenumbers. Each trajectory contained at least several hundred spectra. Figure 2 shows a typical $C(t)$ of a molecule; the trajectory was recorded with an excitation rate of $2.6 \times 10^5 \text{ s}^{-1}$ at 594 nm. The distribution of spectral means (Fig. 2 inset) reflects the transition frequencies accessed by the single molecule in the particular environment. Molecules in different environments have distributions that are significantly different in centre positions (620–680 nm) and in widths (full-widths at half-maximum, 4–20 nm).

The autocorrelation function has a spike at zero time (Fig. 2), due to uncorrelated measurement noise for spectral means (see Fig. 1b for error bars) and spectral fluctuations faster than the time resolution. $C(t)$ can be fitted by a double exponential decay with the spike at zero time:

$$C(0) = \sigma_0^2 + \sigma_1^2 + \sigma_2^2 \quad \text{for } t = 0 \quad (1a)$$

$$C(t) = \sigma_1^2 \exp(-k_1 t) + \sigma_2^2 \exp(-k_2 t) \quad \text{for } t > 0 \quad (1b)$$

The dashed line in Fig. 2 is a fit with a spike, $\sigma_0^2 = 2,000 \text{ cm}^{-2}$; a fast component, $k_1 = 1.9 \text{ s}^{-1}$ and $\sigma_1^2 = 5,000 \text{ cm}^{-2}$; and a slow component, $k_2 = 0.022 \text{ s}^{-1}$ and $\sigma_2^2 = 10,700 \text{ cm}^{-2}$.

To understand the mechanism of these components of spectral fluctuations, we note that molecules do not exhibit noticeable lateral translation beyond a range of a few nanometres. In addition, no detectable variations in the orientation of the transition dipole were observed in this system⁶. Thus we attribute the spectral fluctuations to changes of nuclear coordinates—either the intramolecular coordinates (such as conformations of a side chain) or the intermolecular coordinates (such as hydrogen bonds and collective nuclear coordinates of the interacting environment). The observation that the autocorrelation function of the single molecule has a fast and a slow component implies the existence of at least two quasi-independent distributions of spectral means, with variances of σ_1^2 and σ_2^2 . These two distributions relate to different variations in nuclear coordinates taking place at different rates (k_1 and k_2).

We evaluate the excitation-rate dependence of k_1 and k_2 for many trajectories of many molecules to determine whether the spectral fluctuations are spontaneous or photoinduced. We observed that the fast component (k_1) is independent of excitation rate and excitation wavelength (Fig. 3a), and that the slow

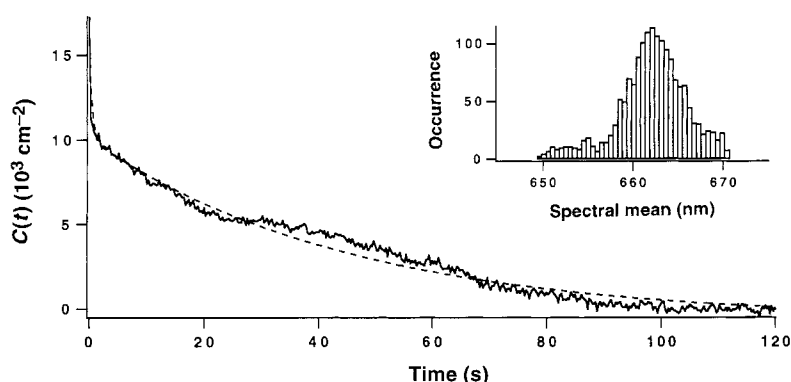


FIG. 2 The autocorrelation function, $C(t) = \langle v(0)v(t) \rangle - \langle v \rangle^2$, derived from the trajectory of spectral means of a single sulphorhodamine 101 molecule at an excitation rate of $2.6 \times 10^5 \text{ s}^{-1}$ at 594 nm. The dashed line is a fit to the data using equation (1b) in the text; parameter values are $k_1 = 1.9 \text{ s}^{-1}$, $k_2 = 0.022 \text{ s}^{-1}$, $\sigma_0^2 = 2,000 \text{ cm}^{-2}$ and $\sigma_2^2 = 10,700 \text{ cm}^{-2}$. Inset, the distribution of the spectral means derived from the same trajectory.

component (k_2) is excitation-rate-dependent and is different for 532-nm (Fig. 3b) and 594-nm excitation wavelengths (Fig. 3c). The excitation rate was varied by adjusting the power of linearly polarized excitation light (0.1–2 μ W) and by choosing different molecular orientations. The excitation rates were determined from the integrated area per spectrum by assuming a fluorescence quantum yield close to unity and an emission collection efficiency of 10%. The amplitude ratio of the fast and slow components (σ_1^2/σ_2^2) is essentially independent of the excitation rate, but varies from molecule to molecule.

The observation that the fast component is independent of the excitation rate indicates that the fast fluctuation is spontaneous (known as spectral diffusion) rather than photoinduced. To further confirm this conclusion, we used a shutter to repeatedly block the excitation light for a period of 'dark time', T_{off} , after each spectrum collection time, T_{on} , and obtained the autocorrelation function, $C(n)$ (n being the index number of spectra) (Fig. 4 inset). Figure 4a shows the $C(n)$ of a molecule with $T_{\text{off}} = 0$; Fig. 4b shows the $C(n)$ for the same molecule with a $T_{\text{off}} = 0.5$ s at an identical excitation rate during T_{on} . If no spontaneous fluctuations occurred during the dark times, the two curves of $C(n)$ would be identical. We observe, however, that the fast component of $C(n)$ disappeared in Fig. 4b, which provides additional evidence of the spontaneous fluctuation during the dark times.

At room temperature, spontaneous fluctuations on the sub-picosecond timescale are responsible for the broad spectral widths (Fig. 1a). However, the spontaneous fluctuation we observed on the rather long timescale (hundreds of milliseconds) would be

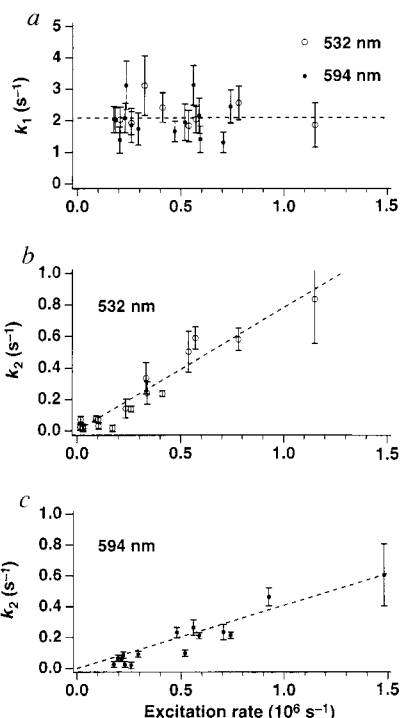


FIG. 3 a, The dependence of k_1 , the fast component of $C(t)$, on the excitation rate. The data points are from different molecules at either 532-nm (open circles) or 594-nm (filled circles) excitation. The independence of k_1 (~ 2.1 s $^{-1}$) on excitation rate and wavelength indicates that the spectral fluctuation is spontaneous. b, The dependence of k_2 , the slow component of $C(t)$, on the excitation rate with 532-nm excitation. The data points are from different molecules. The quasi-linear relationship indicates that the spectral fluctuation is photoinduced. The quantum efficiency for photoinduced spectral change is 7.8×10^{-7} according to the slope. c, The dependence of k_2 on the excitation rate with 594-nm excitation. A quasi-linear relationship is also observed but with a slower photoinduced fluctuation. The quantum efficiency for photoinduced spectral change is 4.1×10^{-7} . The larger error bars at high excitation rates are associated with the shorter trajectories due to photobleaching.

extremely difficult, if not impossible, to detect at room temperature in ensemble-averaged experiments. Although spontaneous fluctuations^{12,13} have been studied for single molecules at cryogenic temperatures, the room-temperature fluctuations we observed have significantly larger magnitude and involve a thermally activated mechanism (rather than tunnelling).

The quasi-linear dependence of k_2 on excitation rate (Fig. 3b, c) clearly indicates the photoinduced nature of the spectral fluctuations at the longer timescale. Although we cannot rule out the possibility of a quadratic dependence on excitation power, multiphoton processes (such as excited state absorption) are not likely to occur at the low power level ($< 10^5$ W cm $^{-2}$) used.

There are two possible mechanisms for the photoinduced spectral fluctuation. First, a molecule can access other ground-state potential minima with different spectra through relaxations from its singlet excited state, a phenomenon known as 'non-photochemical hole burning'¹⁴. Single-molecule 'nonphotochemical hole burning' has been observed at cryogenic temperature^{15–17}, but the spectral shifts are much smaller than those observed in our experiment. Second, the spectral fluctuations can also be associated with radiationless relaxations, probably from the triplet state. The energy released in each relaxation event is quickly dissipated into the substrate, and the original temperature of the system is re-established before the next photoexcitation. However, the energy released can lead to variations of the nuclear coordinates, resulting in a new spectrum in subsequent photoexcitations. At present, we are not able to distinguish experimentally between these two mechanisms. The slow rate of photoinduced fluctuation is associated with the small quantum efficiency for either of the two mechanisms. The apparently steeper slope in Fig. 3b is probably associated with the higher photon energy of the 532-nm excitation.

For either the fast or slow spectral fluctuation component, both thermal and photoinduced contributions are possible. Dominated by the thermal fluctuation, k_1 receives little contribution, if any, from the photoinduced fluctuation because of its small quantum efficiency. Conversely, k_2 has some contribution from the thermal

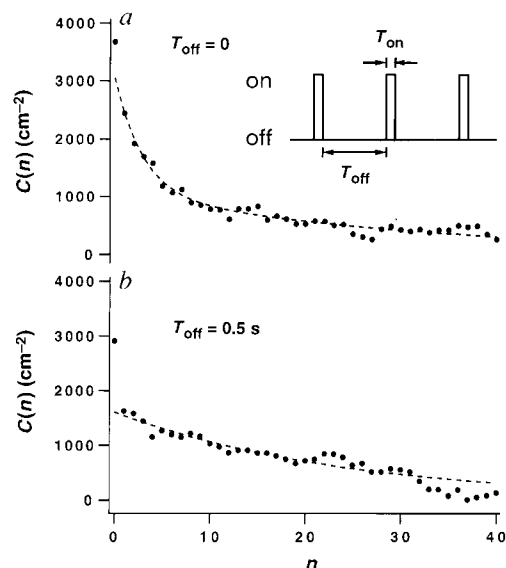


FIG. 4 The inset shows schematically the time dependence of the excitation light; this light is periodically blocked for a dark time, T_{off} , after each spectral collection time, T_{on} (170 ms). a, The autocorrelation function $C(n)$, for a single molecule under continuous 594-nm excitation ($T_{\text{off}} = 0$), n being the index number of the spectra. The dashed line is a double exponential fit, $C(n) = \{2.000 \exp(-0.35n) + 1.050 \exp(-0.03n)\}$ cm $^{-2}$. b, The $C(n)$ for the same molecule taken with a dark time $T_{\text{off}} = 0.5$ s and the same excitation rate during T_{on} . The fast component disappears, proving the spontaneous (thermal) fluctuation during the dark time. The dashed line is a single exponential fit with a decay constant $\gamma = 0.04$.

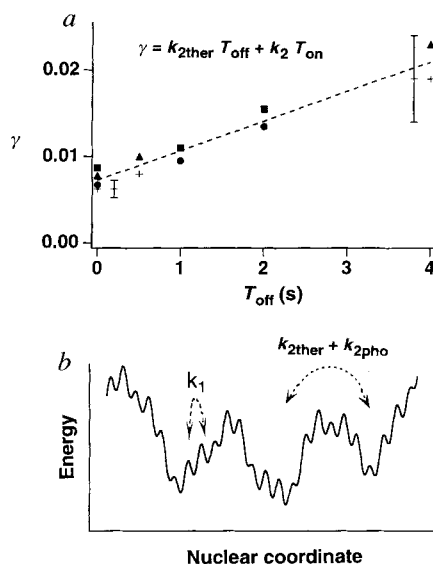


FIG. 5 a, Plot of γ , the decay constant of $C(n)$, as a function of the dark time T_{off} . Data points derived from spectral trajectories of the same sulphorhodamine 101 molecule are labelled with the same symbols. The slope of the dashed line gives the slow thermal rate $k_{2\text{ther}} = 0.02 \pm 0.01 \text{ s}^{-1}$, while the intercept gives the photoinduced rate, $k_{2\text{pho}} = 0.02 \pm 0.01 \text{ s}^{-1}$ at the excitation rate of $2 \times 10^5 \text{ s}^{-1}$ during T_{on} (170 ms). b, Schematic of the potential energy surface of the nuclear coordinates. Our analyses of the spectral trajectories indicate the existence of two distinctly different types of barrier heights. The small barriers are associated with fast thermal fluctuation (k_1), and the large (gross) barriers are associated with the photoinduced ($k_{2\text{pho}}$) and slow thermal ($k_{2\text{ther}}$) fluctuations.

rate ($k_{2\text{ther}}$), but is often dominated by the photoinduced rate ($k_{2\text{pho}} = k_2 - k_{2\text{ther}}$) for a broad range of excitation intensity. Although it is difficult to extrapolate the slower $k_{2\text{ther}}$ at zero excitation rate (Fig. 3b, c), $k_{2\text{ther}}$ can be determined by using the scheme shown in Fig. 4 inset and by evaluating the autocorrelation functions, $C(n)$, at different dark times, T_{off} . When neglecting the fast component k_1 ,

$$C(n) = C(0) \exp\{-nk_{2\text{ther}}(T_{\text{off}} + T_{\text{on}}) - nk_{2\text{pho}}T_{\text{on}}\} \\ = C(0) \exp(-n\gamma) \quad (2)$$

where $\gamma = k_{2\text{ther}}T_{\text{off}} + (k_{2\text{ther}} + k_{2\text{pho}})T_{\text{on}}$ is a single exponential rate constant. Therefore, a longer dark time results in a faster decay of $C(n)$, as is evident by comparing the decay of the curve in Fig. 4b to the slow component of the curve in Fig. 4a. In Fig. 5a, we plot γ as a function of T_{off} for several trajectories of a few molecules at a constant excitation rate during T_{on} . The slope of the fitted line gives $k_{2\text{ther}} = 0.02 \pm 0.01 \text{ s}^{-1}$, whereas the intercept results in $k_{2\text{pho}} = 0.02 \pm 0.01 \text{ s}^{-1}$.

The large k_1 and the small $k_{2\text{ther}}$ indicate that there are two distinctly different types of barrier in the potential-energy surface of the nuclear coordinates. As illustrated schematically in a one-dimensional cross-section of the ground-state potential energy surface (Fig. 5b), the local potential minima connected with the small barriers can be accessed through the fast thermal fluctuation (k_1), whereas the gross potential minima connected with the large barriers can only be accessed through the slow photoinduced fluctuation ($k_{2\text{pho}}$) and the much slower thermal fluctuation ($k_{2\text{ther}}$). Within each gross minimum, there is a spectral mean distribution (with variance of σ_j^2) associated with the different local minima. Independently, there is another spectral mean distribution (with variance of σ_j^2) associated with the different gross minima. The activation energies for k_1 and $k_{2\text{ther}}$ might be evaluated from the temperature dependence of the spectral trajectories.

Although a microscopic description of the nuclear coordinates

awaits further studies, our room-temperature experiments have provided new insights into the general features of the potential-energy surface and the dynamical properties of a single molecule in a particular environment. Similar structures of potential energy surfaces have been put forward for proteins¹⁸, and the kind of analysis described here may allow them to be explored. □

Received 22 May; accepted 18 November 1996.

1. Betzig, E. & Chichester, R. J. *Science* **262**, 1422–1425 (1993).
2. Nie, S. Chiu, D. T. & Zare, R. N. *Science* **266**, 1018–1021 (1994).
3. Macklin, J. J., Trautman, J. K., Harris, T. D. & Brus, L. E. *Science* **272**, 255–258 (1996).
4. Trautman, J. K., Macklin, J. J., Brus, L. E. & Betzig, E. *Nature* **369**, 40–42 (1994).
5. Xie, X. S., Bian, R. X. & Dunn, R. C. in *Focus of Multidimensional Microscopy* Vol. 1 (eds Chen, P. C., Hwang, P. P., Wu, J. L., Wang, G. & Kim, H.) (World Scientific, Teaneck, NJ, in the press).
6. Xie, X. S. & Dunn, R. C. *Science* **265**, 361–364 (1994).
7. Ambrose, W. P., Goodwin, P. M., Martin, J. C. & Keller, R. A. *Science* **265**, 364–367 (1994).
8. Bian, R. X., Dunn, R. C., Xie, X. S. & Leung, P. T. *Phys. Rev. Lett.* **75**, 4772–4775 (1995).
9. Schmidt, Th., Schutz, G. J., Baumgartner, W., Gruber, H. J. & Schindler, H. J. *Phys. Chem.* **99**, 17662–17668 (1995).
10. Bernard, J., Fleury, L., Talon, H. & Orrit, M. J. *Chem. Phys.* **98**, 850–859 (1993).
11. Basche, Th., Kummer, S. & Brauchle, C. *Nature* **373**, 132–134 (1995).
12. Ambrose, P. A. & Moerner, W. E. *Nature* **349**, 225–227 (1991).
13. Reilly, P. D. & Skinner, J. L. J. *Chem. Phys.* **102**, 1540–1552 (1995).
14. Jankowiak, R. & Small, G. J. *Science* **237**, 618–625 (1987).
15. Orrit, M. & Bernard, J. *Phys. Rev. Lett.* **65**, 2716–2719 (1990).
16. Moerner, W. E. et al. *J. Phys. Chem.* **98**, 7382–7389 (1994).
17. Basche, Th., Ambrose, W. P. & Moerner, W. E. *J. Opt. Soc. Am. B* **9**, 829–836 (1992).
18. Frauenfelder, H. & Wolynes, P. G. *Phys. Today* **47**, 58–64 (1994).

ACKNOWLEDGEMENTS. We thank R. Dunn, J. Trautman, P. Barbara, K. Eiseenthal, C. Johnson, S. Colson and G. Holton for discussions. Pacific Northwest National Laboratory is operated for the US Department of Energy (DOE) by Battelle. This work was supported by the Office of Basic Energy Sciences, Chemical Science Division, US DOE.

CORRESPONDENCE should be addressed to X.S.X. (e-mail: xs_xie@pnl.gov).

Silurian hydrothermal-vent community from the southern Urals, Russia

Crispin T. S. Little*, Richard J. Herrington*,
Valeriy V. Maslennikov†, Noel J. Morris*
& Viktor V. Zaykov†

* The Natural History Museum, Cromwell Road, London SW7 5BD, UK
† Institute of Mineralogy, Urals Branch Russian Academy of Sciences,
Miass, Chelyabinsk district 456301, Russia

MODERN hydrothermal-vent communities are remarkable for being dependent on bacterial chemosynthetic primary production and for having a high percentage of endemic taxa (95% at the species level)^{1–3}. Based on phylogenetic analyses, it has been suggested that some of these taxa are Mesozoic or even Palaeozoic relicts, and that the vent environment has thus acted as a refuge against evolutionary pressures, such as mass extinctions, that affect other ecosystems^{1,2,4}. However, little is known about ancient vent communities because fossils have been reported from very few^{5–11} of a thousand or so documented vent deposits¹². Here we describe a macrofossil assemblage of monoplacophoran molluscs, inarticulate brachiopods, vestimentiferan tube-worms and other tubes, probably of polychaete origin, from the Silurian Yaman Kasy deposit¹². The assemblage represents the oldest, and most diverse, fossil hydrothermal-vent community known, and shares vestimentiferan and polychaete tube-worms with both modern vent communities^{1,2} and other ancient vent assemblages^{7–12}, but is unique in having brachiopods and monoplacophorans. Modern vent communities are not refuges for these Silurian shelly vent taxa, a finding that may have implications for the refuge hypothesis.

The Yaman Kasy volcanogenic massive sulphide deposit (Orenburg district, southern Urals, Russia) formed in a Silurian back-arc basin^{13–16}, and can be compared directly to modern sea-floor back-arc vent deposits¹⁷ on the basis of similarities in primary mineralogical textures and palaeo-fluid temperatures (R.J.H.,

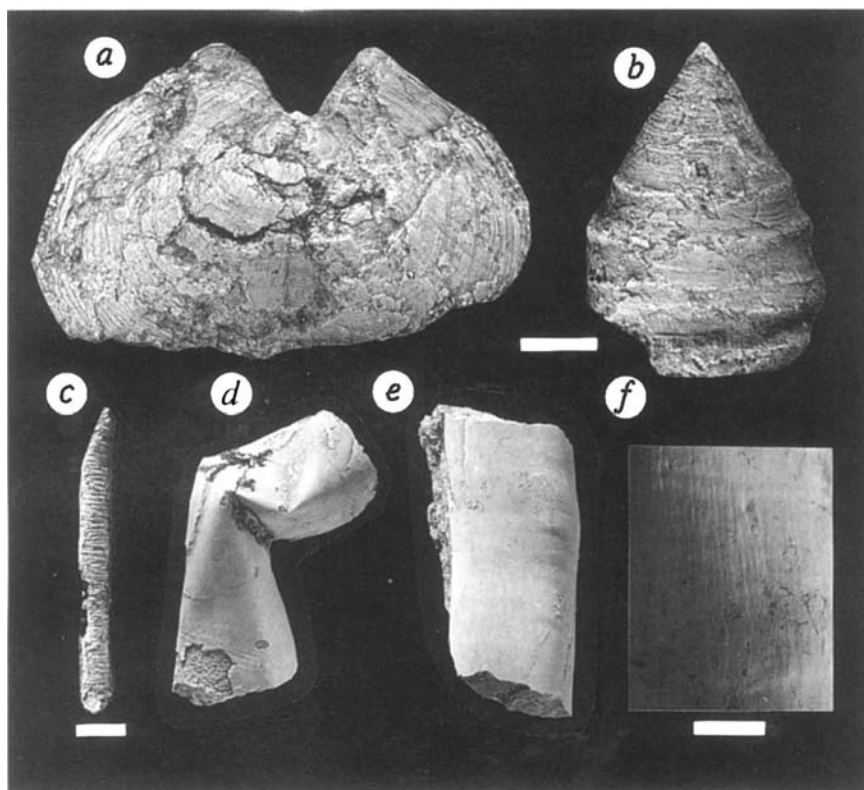


FIG. 1 Representative fossils from the Yaman Kasy deposit. *a*, Acrotretacean inarticulate brachiopod, epoxy-resin external mould of pedicle valve. *b*, Monoplacophoran mollusc, anterior view. This conical species is similar to, but much larger than, several early Palaeozoic monoplacophoran genera from the Urals²⁷. *c*, Presumed polychaete tube internal mould. *d*, *e*, Vestimentiferan tube external morphology. *f*, Detail of vestimentiferan tube wall ornament. Scale bars: *a*, *b*, *d*, *e*, 10 mm; *c*, 3 mm; *f*, 2 mm.

unpublished data). The sulphides near the top of the Yaman Kasy deposit preserve a low-diversity, *in situ* fossil assemblage of vestimentiferan and presumed polychaete tube-worms, and an undescribed species of both acrotretacean inarticulate brachiopod and monoplacophoran mollusc (Fig. 1). (These fossils have been previously ascribed to vestimentiferans, alvinellid-like polychaetes, a *Calypotgena*-like vesicomid bivalve, and an archaeogastropod, respectively¹¹.) The brachiopods (Fig. 1*a*) and monoplacophorans (Fig. 1*b*) are closely associated with the tube-worms, either scattered or in groups of up to ten individuals. At least 24 reasonably complete monoplacophoran specimens are known, although the brachiopod species is rarer, with six disarticulated specimens collected. Both the brachiopods and monoplacophorans are preserved as thin external moulds of pyrite; all original shell material is lacking.

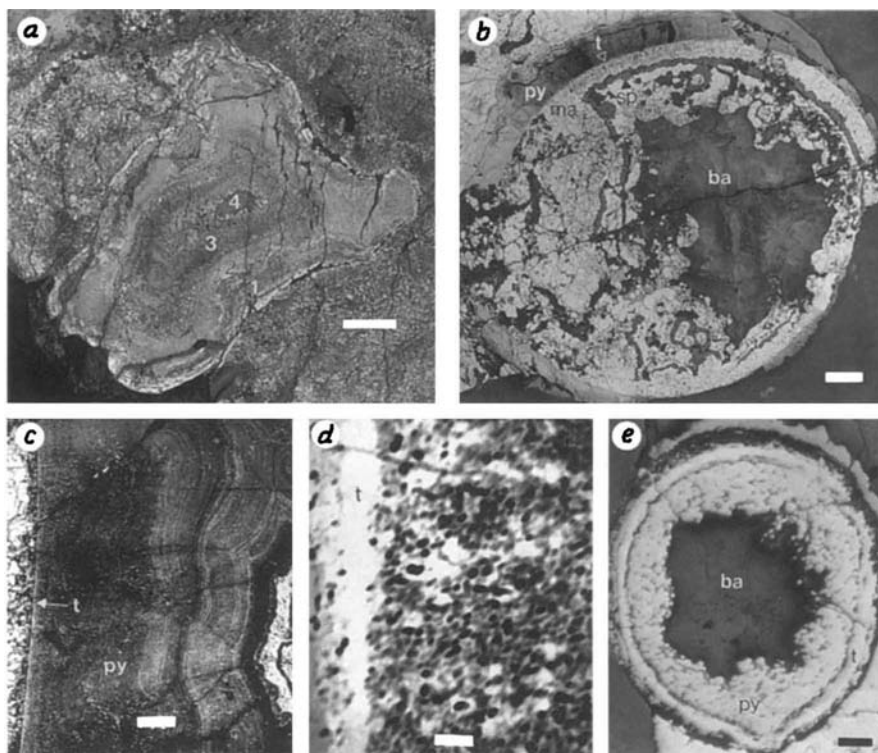
The vestimentiferan and presumed polychaete tubes are common in the deposit (>50 collected specimens of each) and occur as single specimens or discrete clumps of individuals oriented roughly parallel to each other. The presumed polychaete tubes (Fig. 1*c*) are non-tapering, circular in cross-section, 0.8–4.0 mm in diameter, and are a maximum of 75 mm long. The tube walls are formed of pyrite framboids, and are 0.2–1.0 mm thick (Fig. 2*e*). Internal tube moulds show an ornament of cross-cutting annulations (Fig. 1*c*). The vestimentiferan tubes are non-tapering, 10–40 mm in diameter, and are a maximum of 400 mm long (Figs 1*d*, *e* and 2*b*). One specimen shows a conical ending; a few have straight sides and perfectly circular cross-sections; and the rest show varying degrees of pre-diagenetic compaction, being bent (Fig. 1*d*), twisted, and having cross-sections that are elliptical to almost flat. The tube walls are formed by a thin layer of arsenian pyrite, 0.01–0.08 mm thick (Fig. 2*b*), which has an external ornament of wavy longitudinal striations (Fig. 1*f*). Microlaminated pyrite layers, 0.2–2.7 mm thick, are often developed on the exterior and sometimes also the interior of the tube walls (Fig. 2*b*). These layers are porous adjacent to the tube walls (Fig. 2*c*) as a result of circular holes ~1 µm in diameter (Fig. 2*d*). We interpret these to be 'ghosts' of bacterial cells growing on the original organic tube walls of the Yaman Kasy vestimentiferans, because bacterial iron accumulation on vestimentiferan tubes

occurs in modern vent settings^{18,19} and the process may have been similar at ancient sites. Both the Yaman Kasy vestimentiferan and presumed polychaete tubes have zoned mineral infillings (Fig. 2*b*, *e*) that are very like the infillings of vestimentiferan and alvinellid polychaete tubes entombed in recent vent deposits^{7,20,21}. This suggests that worm tubes, both ancient and modern, remained open after initial diagenesis, and acted as conduits for waning hydrothermal temperature flux in the sulphide mounds.

The Yaman Kasy tubes are undoubtedly of biological origin, having quite different wall structures and infillings to black-smoker chimney pieces from the same deposit (Fig. 2*a*). We have identified the larger tubes as vestimentiferans, on the basis of similarities with recent axonobranch vestimentiferans²² in size, gross morphology (flexible, non-tapering tubes with conical closed endings) and life habit (living epifaunally in vent settings). However, more exact determination is impossible because no recent vestimentiferan taxon has an ornament like that of the Yaman Kasy tubes^{22–24}. The presumed polychaete tubes are similar to the annulated small tubes from the Devonian Sibay and Cretaceous Bayda vent assemblages, referred to as alvinellid polychaetes, and vestimentiferans or alvinellids, respectively^{8,11}. Their closest modern analogue may be the annulated sulphide layers that have been observed on the inside of a few *Alvinella* polychaete tubes from the East Pacific Rise (R. Haymon, personal communication).

The Yaman Kasy fossil assemblage represents part of a Silurian hydrothermal-vent community, which we suggest was similar to modern back-arc vent communities in trophic structure^{1,25}, perhaps based on bacterial chemosynthetic primary production with chemosymbiotic associations (vestimentiferans), filter-feeders (brachiopods and presumed polychaetes) and grazers (monoplacophorans; however we note a recent monoplacophoran species with bacterial endosymbionts²⁶). The assemblage shares vestimentiferan and polychaete tube-worms with both modern Pacific vent communities and other ancient vent assemblages, but monoplacophorans and inarticulate brachiopods (representing a class and a phylum, respectively) are unique as vent taxa to Yaman Kasy. It is possible that they are not characteristic elements of other coeval (as yet undiscovered) vent assemblages, because modern vent

FIG. 2 Polished transverse sections of Yaman Kasy tubes and chimney. *a*, Black-smoker vent chimney fragment showing: 1, outer pyrite–chalcopyrite wall; 2, chalcopyrite-rich zone; 3, inner pyrite–chalcopyrite–sphalerite zone; and 4, central sphalerite core. Scale bar, 10 mm. *b*, Vestimentiferan tube showing arsenian pyrite tube wall (t) with externally developed collomorphic laminated pyrite layer (py) and infilling mineralogy of marcasite (m), sphalerite (sp), and barite (ba). Scale bar, 2 mm. The usual sequence of mineral infillings goes inwards from the pyrite tube wall to sulphides (pyrite $\pm$ marcasite $\pm$ sphalerite) to sulphates (barite) $\pm$ silica. *c*, Back-scattered electron micrograph of the pyrite layer of the tube in *b*, showing very fine-scale collomorphic textures throughout and porosity in the basal part. Scale bar, 180 μ m. *d*, Higher magnification of *c* showing that the porosity is caused by circular holes about 1 μ m in diameter. Scale bar, 10 μ m. *e*, Presumed polychaete tube showing tube wall (t) formed of packed pyrite framboids and infilling mineralogy of pyrite framboids (py) and barite plug (ba). Scale bar, 0.1 mm. The usual sequence of mineral zonation goes inwards from a framboidal pyrite tube wall to pyrite to (sphalerite) to barite (or space), but some specimens are completely infilled by framboids, and some have the framboid layer but remain hollow.



communities^{1,2} are spatially heterogeneous in taxonomic structure, and this may also have been the case in the Silurian.

Monoplacophorans and inarticulate brachiopods occur today at very low densities in other deep-sea environments, and were very common in early Palaeozoic non-vent communities. Their absence from modern vent communities shows that the hydrothermal vent environment is not a refuge for the known Silurian shelly vent taxa, and that there has been temporal movement of major groups in and out of the vent ecosystem^{1,2,4}. □

Received 23 October; accepted 14 November 1996.

1. Tunnicliffe, V. *Oceanogr. Mar. Biol. Annu. Rev.* **29**, 319–407 (1991).
2. Tunnicliffe, V. *Palaios* **7**, 338–350 (1992).
3. Karl, D. M. in *The Microbiology of Deep-sea Hydrothermal Vents* (ed. Karl, D. M.) 35–124 (CRC Press, Boca Raton, FL, 1995).
4. Newman, W. A. *Bull. Biol. Soc. Wash.* **6**, 231–242 (1985).
5. Campbell, K. A. & Bottjer, D. J. *Geology* **23**, 321–324 (1995).
6. Boirat, J.-M. & Fouquet, Y. C. R. *Acad. Sci.* **302**, 941–946 (1986).
7. Haymon, R. M., Koski, R. A. & Sinclair, C. *Science* **223**, 1407–1409 (1984).
8. Oudin, E., Bouladon, J. & Paris, J.-P. C. R. *Acad. Sci.* **301**, 157–162 (1985).
9. Oudin, E. & Constantinou, G. *Nature* **308**, 349–353 (1984).
10. Kuznetsov, A. P., Zaykov, V. V. & Maslennikov, V. V. *Izvest. Akad. Nauk Ser. Biol.* **1991**(2), 232–241 (1991).
11. Kuznetsov, A. P., Maslennikov, V. V. & Zaykov, V. V. *Izvest. Akad. Nauk Ser. Biol.* **1993**(4), 525–534 (1993).
12. Franklin, J. M., Sangster, D. M. & Lydon, J. W. *Econ. Geol.* (75th Ann. Vol.), 485–627 (1981).
13. Zaykov, V. V. *Volcanicity and Massive Sulphide Hills of Palaeozoic Margins (Urals and Siberian Massive Sulphide-bearing Zones for example)* (Nauka, Moscow, 1991).
14. Zonenshain, L. P., Kuzmin, M. I. & Natapov, L. M. *Geology of the USSR: A Plate Tectonic Synthesis* (Geodynam. Ser. 21, Am. Geophys. Union, Washington DC, 1990).
15. Zonenshain, L. P. et al. *Tectonophysics* **109**, 95–135 (1984).
16. Koroteev, V. A. *Tectonophysics* **127**, 361–369 (1986).
17. Ishibashi, J. & Urabe, T. in *Backarc Basins: Tectonic and Magmatism* (ed. Taylor, B.) 451–495 (Plenum, New York, 1995).
18. Tunnicliffe, V. & Fontaine, A. R. *J. Geophys. Res.* **92**, 11303–11314 (1987).
19. Juniper, S. K. & Tebo, B. M. in *The Microbiology of Deep-sea Hydrothermal Vents* (ed. Karl, D. M.) 219–247 (CRC Press, Boca Raton, FL, 1995).
20. Haymon, R. M. & Kastner, M. *Earth Planet. Sci. Lett.* **53**, 363–381 (1981).
21. Cook, T. L. & Stakes, D. S. *Science* **267**, 1975–1979 (1995).
22. Jones, M. L. *Bull. Biol. Soc. Wash.* **6**, 117–1158 (1985).
23. Southward, E. C. *J. Nat. Hist.* **25**, 859–881 (1991).
24. Southward, E. C., Tunnicliffe, V. & Black, M. *Can. J. Zool.* **73**, 282–295 (1995).
25. Desbruyères, D., Alayse-Danet, A.-M., Ohta, S. & Scientific Parties of BIOLAU and STARMER *Crust. Mar. Geol.* **116**, 227–243 (1994).
26. Haszprunar, G., Schaefer, K., Warén, A. & Hain, S. *Phil. Trans. R. Soc. Lond. B* **347**, 181–185 (1995).
27. Doguzhaeva, L. A. *Paleont. J.* **1**, 20–23 (1972).

ACKNOWLEDGEMENTS: We thank the MMCK company for access to the Yaman Kasy pit.

CORRESPONDENCE and requests for materials should be addressed to C.T.S.L. (e-mail: c.little@nhm.ac.uk).

Respiration rates in bacteria exceed phytoplankton production in unproductive aquatic systems

Paul A. del Giorgio*, Jonathan J. Cole* & André Cimleris†

* Institute of Ecosystem Studies, Box AB, Millbrook, New York 12545-0129, USA

† Department of Biology, McGill University, 1205 Dr Penfield, Montreal, Quebec H3A 1B1, Canada

PLANKTONIC bacteria are a fundamental component of the organic carbon cycle in aquatic systems¹. Organic carbon consumption by planktonic bacteria is the sum of bacterial production (BP) and bacterial respiration (BR). It is now estimated that 30–60% of phytoplankton production (the amount of inorganic carbon fixed by phytoplankton photosynthesis, corrected for phytoplankton respiration) in marine and freshwater systems is processed by bacteria^{1–3}. These estimates of carbon flow through bacteria are conservative, however, because losses due to bacterial respiration are seldom directly measured^{4,5}. We report here that bacterial respiration is generally high, and tends to exceed phytoplankton net production in unproductive systems (less than 70 to 120 μ g carbon per litre per day). A large proportion of the world's aquatic systems have phytoplankton productivities below this value⁶. Bacterial growth efficiency (BGE) is the result of BP and BR [BGE = BP/(BR + BP)]. Comparisons of our models of bacterial respiration with published models of bacterial secondary production^{1,7} show that bacterial growth efficiency must range from less than 10% to 25% in most freshwater and marine systems, well below the values commonly assumed in many current ecological models^{1,2,8,9}. The imbalance between

bacterial respiration and phytoplankton production suggests that in unproductive aquatic systems, the biological system is a net source of CO₂.

Bacterial production is routinely measured in aquatic studies, but bacterial respiration is not⁵. Instead, most current ecological models of aquatic carbon flow assume bacterial growth efficiencies in the range of 40 to 60%^{1,2,8–11}, mostly on the basis of the uptake and efficiency of conversion of simple ¹⁴C-labelled organic compounds¹². There is now indication that growth efficiencies of bacteria using natural substrates are substantially lower than these assumed values^{5,13,14}, so ecological models may greatly underestimate the total amount of carbon flowing through bacterio-plankton. We surveyed the literature for direct measurements of planktonic bacterial respiration from marine, estuarine and freshwater systems, and assessed how these rates varied with increasing bacterial abundance and planktonic primary production, with a twofold purpose. First, we compared organic carbon consumed through bacterial respiration to organic supplied by net primary production (NPP); if bacterial respiration exceeds net primary production, bacteria must use external sources of organic matter, and the system is net heterotrophic. Second, we estimated bacterial growth efficiencies and assessed how these vary across aquatic systems. A better knowledge of bacterial growth efficiencies is needed to improve our estimates of carbon flow through aquatic microbial food webs.

Using the data averaged per site and study, we observed a strong relation between bacterial abundance and bacterial respiration, lending credibility to this collection of *in situ* measurements of respiration (Fig. 1a). Although there is much scatter in this relation, both bacterial abundance and respiration varied by three orders of magnitude, so the log-log least-squares slope (o.l.s.) was not significantly different from unity (Table 1). The structural (r.m.a.) slope was significantly higher than unity (Table 1). Bacterial respiration was also positively related to net primary production across a wide range of primary productivities, from ultraoligotrophic oceans to highly eutrophic lakes and estuaries (Fig. 1b). Bacterial respiration was almost two orders of magnitude less variable than net primary production among systems, and both the o.l.s. and r.m.a. slopes of this log-log relationship were significantly lower than unity (Table 1), indicating that increases in net primary production are followed by proportionately smaller increases in bacterial respiration. The analysis of the individual data rather than the means revealed essentially the same patterns in bacterial respiration and is not included here.

Bacterial abundance has been shown to be remarkably constant among aquatic systems^{1–3}, and our data show that planktonic bacterial respiration is at least as uniform as bacterial abundance. Interestingly, 51% of our average observations (and 63% of the individual points) had NPP/BR < 1, including marine, estuarine and freshwater data, and none of these cases occurred with NPP > 120 $\mu\text{g C l}^{-1} \text{ d}^{-1}$ (Fig. 1b). Bacterial respiration thus tended to exceed net primary production in aquatic systems with net primary production below 100 $\mu\text{g C l}^{-1} \text{ d}^{-1}$, suggesting that the degree of heterotrophy varies systematically with system productivity. This conclusion is further supported by the strong positive correlation

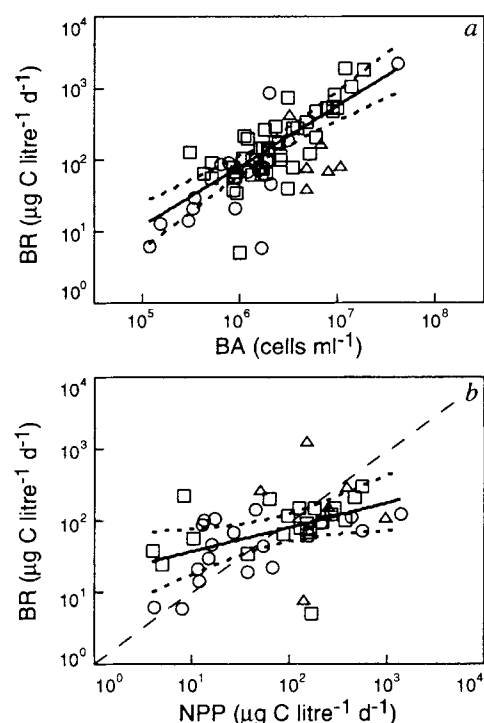


FIG. 1 Planktonic bacterial respiration (BR) as a function of a, bacterial abundance (BA), and b, net primary production (NPP). Symbols represent data from marine (circles), freshwater (squares) and estuarine (triangles) systems. Plots show data averaged per site for each study, and the solid line is the least-square regression with 95% confidence intervals. The dashed line represents equality. The least-squares regression and structural regression parameters for both relationships are shown in Table 1.

found between the NPP/BR ratio and both chlorophyll concentration and phytoplankton biomass in two subsets of the data set (not shown). A large proportion of the world's aquatic systems have primary productivities below 100–120 $\mu\text{g C l}^{-1} \text{ d}^{-1}$ (ref. 6).

Can we confidently conclude that bacterial respiration exceeds phytoplankton production in many unproductive aquatic systems, based on these heterogeneous estimates of bacterial respiration and primary production from the literature? The conclusions of the analysis are similar irrespective of whether o.l.s. or r.m.a. slopes are considered (Table 1), and an unrealistically high error (coefficient of variation >250%) in the independent variables would be required to modify these conclusions. Most of the primary production data are measures of ¹⁴C uptake made using standard techniques within the past ten years, so these data should be reliable, but the bacterial respiration data are more difficult to judge. One way to assess the overall validity of the bacterial respiration data is to determine whether the resulting growth efficiencies are within reasonable ranges. We calculated bacterial

TABLE 1 Parameters of linear regression models relating bacterial respiration to bacterial abundance and net primary production

Model	X	Y	Intercept	Slope	r ²	s.e.	N	CF
o.l.s.	BA	BR	-3.51 ± 1.26	0.67 < 0.88 < 1.09	0.53	0.41	72	1.56
r.m.a.	BA	BR	-5.21	0.98 < 1.15 < 1.35	0.53		72	
o.l.s.*	NPP	BR	1.21 ± 0.38	0.16 < 0.34 < 0.52	0.30	0.43	47	1.63
r.m.a.	NPP	BR	0.86	0.48 < 0.62 < 0.78	0.30		47	

Parameters of linear regression models, of the form $\log Y = a(\log X) - b$, relating bacterial respiration (BR, $\mu\text{g C l}^{-1} \text{ d}^{-1}$) to bacterial abundance (BA, $\mu\text{g C l}^{-1} \text{ d}^{-1}$) and net primary production (NPP, $\mu\text{g C l}^{-1} \text{ d}^{-1}$). Ordinary least squares (o.l.s.) and reduced major axis (r.m.a.) models are provided, for the averages per site and study (means), together with the 95% confidence intervals for the regression parameters, the standard error of the estimate (s.e.), significance of the regression (P) and a factor (CF) to correct for bias introduced by log transformation⁷. All the equations are highly significant ($P < 0.001$). Confidence limits for the r.m.a. slope allow for the asymmetry of the distribution of the r.m.a. estimator³⁰.

* This is equation 1 (see text) used to estimate BK from NPP.

growth efficiencies by combining our bacterial respiration data with estimates of bacterial production for each site (Fig. 2). Roughly similar ranges and median values of bacterial growth efficiencies for each type of system were obtained when bacterial production was estimated from bacterial abundance or from net primary production, suggesting that the patterns in bacterial growth efficiency shown in Fig. 2 are robust.

Bacterial growth efficiencies ranged from less than 5% to more than 60% across systems (Fig. 2). The median BGE for marine sites was 0.24 (0.23s.d.) when BP was calculated from BA, and 0.20 (0.17s.d.) when BP was calculated from NPP, and for lakes the medians were 0.17 (0.12s.d.) and 0.23 (0.17s.d.), respectively (Fig. 2). For estuaries, the median BGE calculated based on NPP was 0.21 (0.27s.d.), and much higher when calculated based on BA (0.46, 0.21s.c.). Except for this last estimate, all these values are in good agreement with direct measurements of growth efficiency of bacterial assemblages utilizing natural dissolved organic matter^{13–18}, suggesting that our data on bacterial respiration are realistic. Furthermore, these results support previous reports that bacterial growth efficiencies in aquatic food webs are generally lower than normally assumed in many current ecological models^{1,2,8,9}. One important consequence is that many contemporary models of organic carbon cycling for both freshwater and marine aquatic systems do not balance if the median bacterial growth efficiencies reported here are applied.

How do these patterns in bacterial respiration among systems relate to published models of bacterial secondary production? Comparisons of the above models of bacterial respiration with existing large-scale empirical relationships^{1,7} further show that bacterial respiration increases far more slowly than bacterial production along gradients of both bacterial abundance⁷ and primary production¹ in freshwater and marine systems. We simulated bacterial growth efficiency along a gradient of net primary production, by calculating bacterial respiration ($\mu\text{g C l}^{-1} \text{ d}^{-1}$) from NPP ($\mu\text{g C l}^{-1} \text{ d}^{-1}$) using equation (1) in Table 1, and bacterial production ($\mu\text{g C l}^{-1} \text{ d}^{-1}$) from NPP using the model of Cole *et al.*¹. For each NPP point, BGE was then calculated as $\text{BP}/(\text{BP} + \text{BR})$. This calculation suggests that bacterial growth efficiency increases as a power function of NPP ($\text{BGE} = 0.02 \pm \text{NPP}^{0.41}$).

Although there is considerable uncertainty associated with these large-scale empirical models, an interesting consequence of the diverging trends in bacterial production and respiration is that there appears to be a general trend of increasing bacterial growth efficiencies along gradients of system enrichment, from <10% in oligotrophic sites to a plateau at ~40% in the most productive systems. Similar trends have been reported in BGE for specific sites^{17,18}, and the increase in BGE with system productivity may be linked to systematic changes in both the rate of supply and nutritional quality of organic carbon substrates available for bacteria^{14,18}. The availability of organic carbon substrates, however, is only one of the factors regulating growth efficiencies: for example, low overall bacterial growth efficiencies in open oceans may be linked to low iron availability¹⁹.

Our results support the observations that many unproductive lakes and estuaries are net heterotrophic systems^{20,21}, where the total carbon processed by bacteria exceeds the carbon fixed by phytoplankton. Allochthonous organic carbon of terrestrial origin is likely to play a major role in this metabolic imbalance. Lakes, for example, are often supersaturated with CO_2 (refs 22, 23), and bacterial degradation of allochthonous organic carbon is thought to be one of the main causes for high CO_2 partial pressure²⁰. Likewise, estuaries and coastal oceans are often heavily subsidized with organic carbon from land and wetland ecosystems^{21,24,25}.

Our results also concur with microbial observations for the open ocean that suggest a large role for bacteria in the organic carbon flow^{8,11}, and further indicate that parts of the ocean are net heterotrophic, at least during certain periods. Methodological uncertainties may be partly responsible for these patterns, but our study only deals with volumetric rates of bacterial respiration and phytoplankton production measured either in the mixed layer or,

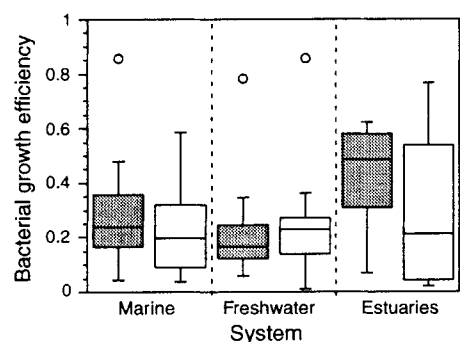


FIG. 2 Box-and-whisker plots showing the range and the median bacterial growth efficiency (BGE) by system. BGE was calculated using the measured rates of bacterial respiration culled from the literature, and rates of bacterial production estimated either from the accompanying bacterial abundance data using the model in ref. 7 (grey boxes) or from net primary production data using the model in ref. 1 (white boxes).

more often, in the euphotic zone of the water column. Bacterial activity in the deep waters of the ocean is low compared to that in surface waters², but photosynthesis is effectively zero, so if these rates were integrated over the entire water column, the imbalance between bacterial respiration and phytoplankton production would be much greater than shown here.

The processes underlying the apparent net heterotrophy in oceans may be different from those in lakes and estuaries. It is problematic to invoke the influence of allochthonous organic matter to fuel excess heterotrophic activity, particularly in the central gyres of the oceans⁴. Organic carbon of terrestrial origin is a significant fraction of dissolved organic carbon (DOC) in marine coastal and shelf waters²⁶, but in open oceans only about 10% of the DOC is now estimated to be of terrestrial origin²⁷. Much of this carbon may not even be biologically available, although photochemical breakdown of DOC may facilitate the utilization by bacteria of recalcitrant humic compounds²⁸. Export of DOC from more productive coastal and upwelling areas to open oceans has also been proposed²⁵. Also, our analysis may mask a succession of periods of net autotrophy in the open ocean, owing to time lags between peaks of primary production in the spring and heterotrophic utilization of the resulting organic carbon later on in the year²⁹. The challenge will be to quantify the relative importance of these alternative processes to net heterotrophy in aquatic systems. □

Methods

Data collection. The relationship between bacterial respiration and bacterial abundance (Fig. 1a) was constructed with data from 20 published articles, and reports of several measurements for the same site were averaged (Fig. 1a) or analysed individually (data not shown). A total of 153 individual data points collected for the literature, which resulted in 62 average values, 20 for open ocean and coastal marine sites, 7 for estuaries and 35 lakes worldwide. Additional data from 10 southern Québec lakes were included (A.C., unpublished results). Most studies reported bacterial respiration measured as oxygen consumption in water filtered through small-pore-size filters (0.8–2 μm) and bacterial abundance measured with epifluorescence counts. Bacterial respiration reported as oxygen consumption per unit time were converted to $\mu\text{g C l}^{-1} \text{ d}^{-1}$, assuming a respiratory quotient (RQ) of 1 (by mol), and that respiration per day was 24 times the hourly rate. All data, except temperature, were log-transformed to attain homoscedasticity and normality. For a subset of the data ($n = 38$), temperature was reported together with BR and BA and we could explore the effect of temperature on the rates of bacterial respiration (data not shown). The following least-squares multiple regression model of bacterial respiration as a function of bacterial abundance and temperature (T) explained 82% of the variance in BR in this subset of the data: $\log \text{BR} = -3.67 (\pm 2.30) + 0.75 (\pm 0.35) \log \text{BA} + 0.059 (\pm 0.01) T$; $r^2 = 0.82$, $n = 38$, s.e. = 39, $P < 0.001$. The temperature coefficient for BR is not significantly different for the one reported by White *et al.*⁷ for bacterial production, suggesting that temperature dependence of bacterial growth efficiencies is probably weak and should not influence the patterns in BGE among systems discussed here.

The relation between BR and NPP (Fig. 1b) was constructed with data extracted from 17 published articles, and reports of several measurements for the same site were averaged. A total of 81 individual data points collected from the literature resulted in 47 average values, 18 for open ocean and coastal marine sites, 7 for estuaries and 22 for lakes worldwide. Volumetric rates of net primary production per unit time were converted to $\mu\text{g C litre}^{-1} \text{ d}^{-1}$, assuming 12 times the hourly rate. The complete data sets are available from the authors or from the Repository of Unpublished Data, CISTI, National Research Council of Canada, Ottawa, Ontario K1A 0S2.

Statistical analysis. All the data were analysed using ordinary least squares (o.l.s.) regression, for comparison with previously published empirical models. The o.l.s. slope will systematically underestimate the true slope, however, owing to error in both the dependent and the independent variables, so we also calculated the reduced major axis (r.m.a.) slopes for the structural relationship between BR–BA and BR–NPP³⁰. The complete o.l.s. and r.m.a. regression parameters are shown in Table 1.

Calculation of BGE. For Fig. 2, bacterial growth efficiencies were calculated as $\text{BGE} = \text{BP}/(\text{BP} - \text{BR})$, where BR are bacterial respiration data collected from the literature and used in Fig. 1a, b. For each BR point, bacterial production (BP, in $\mu\text{g C l}^{-1} \text{ d}^{-1}$) was calculated from either the accompanying net primary production data in Fig. 1b using the equation of Cole *et al.*¹, or from the accompanying bacterial abundance data in Fig. 1a using the equation of White *et al.*⁷. Data are presented as box-and-whisker plots showing the median and range of BGE for each system for the two methods for estimating bacterial production.

Received 15 July; accepted 5 November 1996.

1. Cole, J. J., Findlay, S. & Pace, M. *Mar. Ecol. Prog. Ser.* **43**, 1–10 (1988).
2. Williams, P. J. leB. *Kiel. Meeresforsch.* **51**, 1–28 (1981).
3. Ducklow, H. W. & Carlson, C. A. *Adv. Microb. Ecol.* **12**, 113–181 (1992).

4. Hansell, D. A., Bates, N. R. & Gundersen, K. *Mar. Chem.* **51**, 201–202 (1995).
5. Jahnke, R. A. & Craven, D. B. *Limnol. Oceanogr.* **40**, 436–441 (1995).
6. Lieth, H. & Whittaker, R. H. *Primary Production of the Biosphere* (Springer, New York, 1975).
7. White, P. A., Kalf, J., Rasmussen, J. B. & Gasol, J. M. *Microb. Ecol.* **21**, 99–118 (1991).
8. Cho, B. C. & Azam, F. *Nature* **332**, 441–443 (1988).
9. Strayer, D. *Limnol. Oceanogr.* **33**, 1217–1220 (1988).
10. Wright, R. T. *Hydrobiologia* **159**, 111–117 (1988).
11. Fuhrman, J. A., Sleeter, T. D., Carlson, C. A. & Proctor, L. M. *Mar. Ecol. Prog. Ser.* **57**, 207–217 (1989).
12. Crawford, C. C., Hobbie, J. E. & Webb, K. L. *Ecology* **55**, 551–563 (1974).
13. Kirchman, D. L., Suzuki, Y., Garside, C. & Ducklow, H. W. *Nature* **352**, 612–614 (1991).
14. Kroer, N. *Limnol. Oceanogr.* **38**, 1282–1290 (1993).
15. Danieles, G., Riemann, B. & Williams, P. J. leB. *J. Plankton Res.* **16**, 105–113 (1994).
16. Kristensen, K., Nielsen, H., Riemann, B. & Fuhrman, J. A. *Microb. Ecol.* **24**, 145–160 (1992).
17. Biddanda, B., Opsahl, S. & Benner, R. *Limnol. Oceanogr.* **39**, 1259–1275 (1994).
18. Griffith, P. C., Douglas, D. J. & Wainright, S. C. *Mar. Ecol. Prog. Ser.* **59**, 263–270 (1990).
19. Tortell, P. D., Maldonado, M. T. & Price, N. M. *Nature* **383**, 330–332 (1996).
20. del Giorgio, P. A. & Peters, R. H. *Limnol. Oceanogr.* **39**, 772–787 (1994).
21. Findlay, S., Pace, M. L., Lints, D. & Howe, K. *Mar. Ecol. Prog. Ser.* **89**, 147–153 (1992).
22. Cole, J. J., Caraco, N. F., Kling, G. W. & Kratz, T. K. *Science* **265**, 1568–1570 (1994).
23. Kling, G. W., Kipphut, G. W. & Miller, M. C. *Science* **251**, 298–301 (1990).
24. Hopkinson, C. S. *Mar. Biol.* **87**, 19–32 (1985).
25. Smith, S. V. & Mackenzie, F. T. *Global Biogeochem. Cycles* **1**, 187–198 (1987).
26. Moran, M. A., Pomeroy, L. R., Sheppard, E. S., Atkinson, L. P. & Hodson, R. E. *Limnol. Oceanogr.* **36**, 1134–1149 (1991).
27. Meyers-Schulte, K. J. & Hedges, J. I. *Nature* **321**, 61–63 (1986).
28. Kieber, D. J., McDaniel, J. & Mopper, K. *Nature* **341**, 637–639 (1989).
29. Ducklow, H. W., Kirchman, D. L., Quinby, H. L., Carlson, C. A. & Dam, H. G. *Deep-Sea Res.* **40**, 245–263 (1993).
30. McArdle, B. H. *Can. J. Zool.* **66**, 2329–2339 (1987).

ACKNOWLEDGEMENTS. We thank Y. Prairie, J. Kalf, M. Pace, R. Peters and J. Rasmussen for critical comments on the manuscript, and T. H. S. Canut for encouragement. P.A.D.G. was supported by a Canadian NSERC postdoctoral fellowship, and A.C. was supported by FURNAS and CAPES (Brazil); J.J.C. was supported in part by a grant from the National Science Foundation. This is a contribution to the Institute of Ecosystem studies.

CORRESPONDENCE and requests for materials should be addressed to P.A.D.G. (e-mail: c1434@er.uqam.ca).

Episodic adaptive evolution of primate lysozymes

Walter Messier & Caro-Beth Stewart

Department of Biological Sciences, University at Albany, State University of New York, Albany, New York 12222, USA

ALTHOUGH the darwinian concept of adaptation was established nearly a century ago, it has been difficult to demonstrate rigorously that the amino-acid differences between homologous proteins from different species have adaptive significance. There are currently two major types of sequence tests for positive darwinian selection on proteins from different species: sequence convergence, and neutral rate violation (reviewed in ref. 1). Lysozymes from the stomachs of cows and langur monkeys, two mammalian species displaying fermentation in the foregut, are an example^{2,3} of amino-acid sequence convergence among homologous proteins^{4–6}. Here we combine tests of neutral rate violation with reconstruction of ancestral sequences to document an episode of positive selection on the lineage leading to the common ancestor of the foregut-fermenting colobine monkeys. This analysis also detected a previously unsuspected adaptive episode on the lineage leading to the common ancestor of the modern hominoid lysozymes. Both adaptive episodes were followed by episodes of negative selection. Thus this approach can detect adaptive and purifying episodes, and localize them to specific lineages during protein evolution.

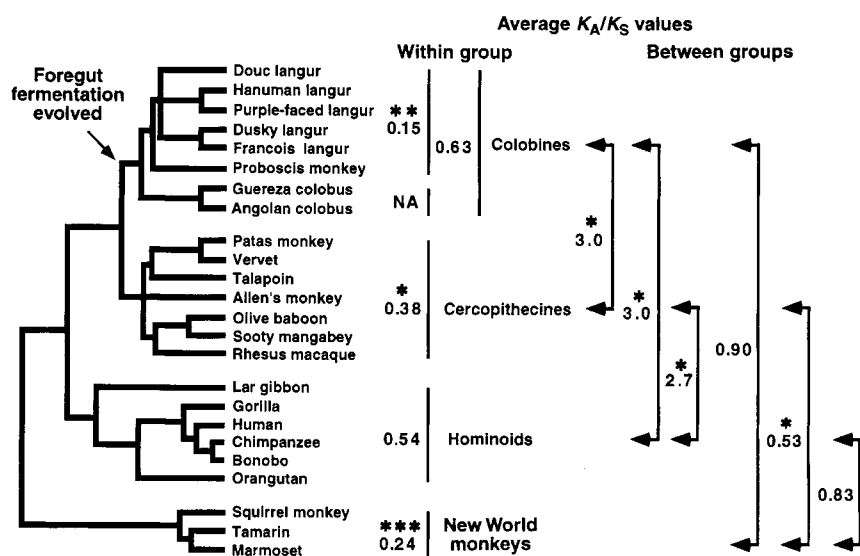
The colobine Old World monkeys are unique among the primates in having a complex foregut in which bacteria ferment leafy plant materials⁷, followed by a true stomach that expresses high levels of the bacteriolytic enzyme, lysozyme^{4,6,8}. Other primates, including the cercopithecine Old World monkeys, have simple stomachs⁷ with lysozyme expressed only in the pyloric region⁶. The cercopithecines, therefore, can serve as closely related 'negative controls' for the comparative study of the evolution of lysozyme in the colobines. Using the phylogeny of the primates (Fig. 1) as a guide, we sequenced the protein-coding

region of the lysozyme gene from species representing each of the major lineages of catarrhines (colobines, cercopithecines and hominoids), as well as from some New World monkeys for primate outgroup comparisons (see Supplementary Information). Visual analysis of these sequences shows that there are many non-synonymous nucleotide differences (that is, differences that would cause amino-acid replacements) compared with the number of synonymous nucleotide differences (those that would not cause amino-acid replacements); this suggests that the proteins might have evolved under positive darwinian selection.

Rigorous demonstration that a protein has evolved more rapidly than the neutral substitution rate requires more than simply counting non-synonymous and synonymous differences between sequences. Proper consideration of the genetic code requires a comparison of the number of non-synonymous substitutions per non-synonymous site (termed K_A in the method^{9,10} used here) with the number of synonymous substitutions per synonymous site (termed K_S) between pairs of homologous protein-coding genes, corrected for multiple hits^{9–11}. Such comparisons are conventionally presented as ratios between the K_A and K_S values, although statistical significance for such comparisons must be calculated using the differences between the values, a convention we follow here. K_A/K_S ratios less than 1.0 are generally taken as evidence that the proteins have evolved under negative or purifying selection; most pairwise comparisons between extant genes show this pattern^{11–15}. Conversely, K_A/K_S ratios significantly greater than 1.0 are considered to be strong evidence of positive selection for amino-acid replacements, because they imply that the rate of non-synonymous substitution exceeds that which can be explained solely by the neutral substitution rate at the locus under study. Significantly elevated K_A/K_S ratios have been documented for very few whole proteins^{13–15}, although domains of certain proteins display this phenomenon^{1,12,16}.

K_A and K_S calculations were made for all possible pairwise comparisons of the primate lysozyme DNA sequences (see Supplementary Information). As predicted by previous studies on lysozyme in the Hanuman langur^{4–6}, nearly all pairwise comparisons between the sequences from the colobines and the cercopithecines showed K_A/K_S ratios greater than 1.0 (range 0.77–3.68). Many of these comparisons are statistically significant, and are

FIG. 1 Phylogeny of the primate species used in this study, showing average K_A/K_S ratios within and between groups. K_A and K_S group averages were calculated as described¹⁷ using the SEND computer program. The within-group average from the two African colobus species could not be calculated because the two sequences are identical. Statistical significance was calculated by t-tests, with P values as follows: *, <0.05; **, <0.01; ***, <0.005. The branching order of the major groups of anthropoid primates (New World monkeys, hominoids and Old World monkeys) has been established by numerous studies. The well-supported mitochondrial DNA phylogeny of the hominoids is presented. The cercopithecine phylogeny represents a combination of molecular and morphological studies. The overall colobine phylogeny is consistent with both our mtDNA phylogeny of this group (R.V. Collura and C.-B.S., manuscript in preparation) and phylogenetic analysis of the complete coding region of the lysozyme sequences (data not shown). Those branching orders shown as multifurcations were not resolved by available sequence data; however, changing the branching



order of these lineages did not change the results of our analyses in any meaningful manner, because the DNA sequences from the species involved are so similar.

among the highest known for any whole protein^{1,12,13}. The large range in these ratios reflects the stochastic nature of the synonymous nucleotide-substitution process¹⁷, as K_S was small and varied more widely than K_A in these pairwise comparisons. Because all colobines are equally related to all cercopithecines, having diverged from a common ancestor about 15 Myr ago¹⁸, comparisons between group averages would be statistically more meaningful than individual comparisons of K_A and K_S ¹⁶. We therefore calculated average K_A and K_S values both within and between phylogenetic groups of these primate lysozymes; K_A/K_S ratios for these group averages are summarized in an evolutionary context in Fig. 1. The between-group average K_A/K_S ratio for all colobine lysozymes versus all cercopithecine lysozymes is statistically significant ($K_A/K_S = 3.0$; $P < 0.05$). Similarly, all comparisons between the colobine and the hominoid lysozymes gave K_A/K_S ratios greater than 1.0 (range 1.33–3.49) and the average is statistically significant ($K_A/K_S = 3.0$; $P > 0.05$). We had expected that comparisons between lysozymes from primates with simple stomachs would yield K_A/K_S ratios less than 1.0, as there was no known change in role for these enzymes, as there had been in the colobine case^{4,6,8}. Contrary to this expectation, all comparisons between the cercopithecine and hominoid sequences yielded elevated K_A/K_S ratios (range 1.02–3.28), with a statistically significant average ($K_A/K_S = 2.7$; $P < 0.05$). These significantly high K_A/K_S values strongly suggest that positive selection for amino-acid replacements occurred during the evolution of catarrhine lysozymes, but cannot identify the lineage.

In contrast to the between-group comparisons, all within-group averages of catarrhine lysozymes show K_A/K_S ratios less than 1.0 (Fig. 1). These within-group comparisons involve sequences that are so closely related that these low ratios suggest that purifying selection has been active during the recent evolution of primate lysozymes. The striking contrast of the purifying selection within groups and the positive selection between groups (Fig. 1) suggests that the adaptive evolution of lysozyme occurred early in catarrhine evolution. If this is the case, then K_A/K_S ratios between the lysozymes from the ancestors of the major lineages should be elevated.

To test this hypothesis, we reconstructed ancestral DNA sequences representing all internal nodes on the tree shown in Fig. 2, using both newly developed maximum-likelihood methods^{19–21} and traditional maximum-parsimony methods^{20,22}. Because the divergence of these primate species was relatively recent, and there are a large number of lysozyme sequences of

appropriate phylogenetic spread, these ancestral sequences were reconstructed with relatively few equally parsimonious alternatives by the maximum-parsimony method, and with generally high probabilities by the maximum-likelihood method. Although maximum-likelihood methods can theoretically reconstruct a different ancestral sequence from that found by maximum-parsimony methods²¹, in all cases examined the maximum-likelihood ancestral lysozyme sequence was identical to one of the maximum-parsimony sequences. The reconstructed DNA sequences representing key primate ancestors (labelled internal nodes in Fig. 2) were then used in pairwise K_A/K_S calculations.

This ancestral analysis of K_A and K_S pinpointed an episode of adaptive lysozyme sequence evolution ($K_A/K_S = 4.7$; $P < 0.05$) on the ancestral colobine lineage (that is, on the colobine lineage after its divergence from the cercopithecine lineage, but before the speciation event that gave rise to the African colobus and the Asian langur clades; between nodes OW and Co in Fig. 2). This period corresponds to the time in evolutionary history when lysozyme was most likely to have been recruited to digest foregut bacteria in the true stomachs of colobine monkeys. In comparison, on the comparable ancestral lineage leading to the cercopithecine monkeys (between nodes OW and Ce in Fig. 2), the K_A/K_S ratio is not significantly different from neutral expectation. This analysis also identified an episode of positive selection ($K_A/K_S = 5.1$; $P < 0.05$) on the ancestral hominoid lineage (between nodes Ca and Ho in Fig. 2) that was not anticipated by previous analyses^{4–6}. We do not have an adaptive explanation for this episode of positive selection, although it might be related to increased neutrophil expression of lysozyme in hominoids as compared with other catarrhines²³. That these episodes of positive selection were followed by periods of purifying selection is supported by the average K_A/K_S ratios of less than 1.0 for all within-group comparisons (Fig. 1). Therefore, combining ancestral sequence reconstruction with pairwise K_A and K_S comparisons allows the detection of specific episodes of positive and negative selection, and localization of these episodes to distinct branches on an evolutionary tree.

Although elevated K_A/K_S ratios are usually taken to suggest that amino-acid replacements are being fixed faster than the neutral substitution rate, it is possible that they might result from a decrease in the fixation of synonymous mutations, rather than an adaptive increase in the fixation of non-synonymous mutations¹. To test for this possibility, we sequenced intron 3 of the lysozyme gene from phylogenetically diverse catarrhines. The

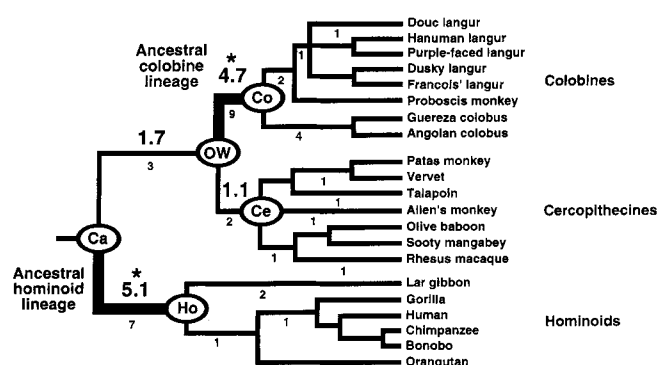


FIG. 2 K_A/K_S ratios between ancestral lysozyme DNA sequences detect early episodes of adaptive molecular evolution. Ancestral sequences (available from C.-B.S. on request) were reconstructed on the full primate phylogeny shown in Fig. 1. K_A/K_S values between ancestral sequences reconstructed by maximum likelihood are shown in a larger typeface above the internal lineages; those ratios with (K_A/K_S) differences that are significant at the 0.05 level are indicated by an asterisk. The minimum number of amino-acid replacements required to explain the sequence differences between these maximum-likelihood reconstructed ancestral sequences are given below the lineages in a smaller typeface. Nodes representing common ancestors of the extant primate groups of interest are labelled as follows: Co, colobine; Ce, cercopithecin; Ho, hominoid; OW, Old World monkey; Ca, catarrhine. Maximum-likelihood ancestral sequences were reconstructed by the marginal and joint methods in PAUP* 4.0 (ref. 20) and by the marginal method in PHYLIP 4.0 (ref. 19), with identical results; these sequences were used for the analyses shown. Ancestral sequences were also inferred by maximum parsimony using the computer programs PAUP* 4.0 (ref. 20) and MacClade 3.0 (ref. 22). For each labelled node there was more than one equally parsimonious ancestral sequence, with the number of ambiguities increasing at earlier nodes; for example, using the primate DNA sequences, there were 2 equally parsimonious ancestral sequences for node Ho, 4 for node Ce, 6 for node Co, 16 for node OW, and 16 for node Ca. The ranges of K_A/K_S values for all equally parsimonious sequences are as follows: Ca to Ho, 3.0–5.9*; Ca to OW, 0.85–3.8; OW to Co, 1.5–5.3*; OW to Ce, 0.56–2.3. Use of a larger data set containing 50 DNA sequences, including non-primate vertebrate lysozymes (both of the conventional chicken and calcium-binding types), gave similar overall results to those from the primate data set, except that the K_A/K_S ratio for the ancestral hominoid lineage was less (maximum-likelihood K_A/K_S = 2.5).

neutral rate of nucleotide substitution calculated for this intron is almost identical to the synonymous substitution rate at this locus. Furthermore, the neutral substitution rate in the lysozyme gene is similar to that of the well-studied η - and δ -globin pseudogenes (W.M. *et al.*, manuscript in preparation). Thus the high K_A/K_S ratios for the catarrhine lysozymes cannot be explained by significant suppression of synonymous substitutions in this gene.

Taken together, the above analyses strongly suggest that there have been two major episodes of positive darwinian selection during the evolution of catarrhine lysozymes. To enumerate and identify the amino-acid replacements that probably occurred during these adaptive episodes, we used maximum-likelihood and maximum-parsimony methods to assign events to specific lineages (Fig. 2). Strikingly, maximum-likelihood analysis suggests that there were nine amino-acid replacements during the brief adaptive episode on the ancestral colobine lineage (Fig. 2), five of which are among those originally identified to have occurred in parallel or convergently along the lineage leading to cow⁴ and other ruminant²⁴ stomach lysozymes. During a similar period along the ancestral cercopithecin lineage, only two amino-acid replacements are reconstructed. Similarly, maximum-likelihood analysis suggests that seven amino-acid replacements occurred during the adaptive episode on the ancestral hominoid lysozyme lineage (Fig. 2). Thus, as suggested by the ancestral K_A/K_S calculations, numerous amino-acid replacements were fixed during the brief adaptive episodes on the ancestral colobine and

hominoid lysozyme lineages. A more complete discussion of these matters will be presented elsewhere.

The evidence for these two adaptive episodes becomes masked when K_A/K_S comparisons are made between lysozymes from more distantly related species. For example, most comparisons of the catarrhine lysozyme sequences with those from the New World monkeys show K_A/K_S ratios less than 1.0 (range 0.38–1.03), and all between-group averages for these comparisons are less than 1.0 (Fig. 1). Similarly, pairwise comparisons between the extant rodent, artiodactyl and primate lysozyme genes produce K_A/K_S ratios significantly less than 1.0 (see Supplementary Information), even though published analyses of artiodactyl lysozymes^{24–26} and our analyses of rodent lysozymes (data not shown) suggest that episodes of positive selection have occurred within both of these lineages. However, these mammalian lysozymes seem to have been mainly under negative selection, so the adaptive episodes are obscured in comparisons involving distantly related sequences. This demonstrates the importance of comparing sequences that diverged within appropriate time frames, if positive darwinian selection is to be detected.

For example, a recent survey of the genetic databases, in which predominantly distantly related sequences were compared, led to the conclusion that positive selection in protein-coding genes is very rare¹². However, this survey failed to identify either mammalian lysozymes or abalone fertilization proteins^{14,15}, although both of these gene families show significantly elevated K_A/K_S ratios when pairwise comparisons are made in a proper phylogenetic context. Thus the survey probably missed many other cases of positive selection, particularly if most proteins evolve in an episodic manner, as has been proposed². The phylogenetic approach we used in comparing K_A and K_S values is capable of teasing apart adaptive and purifying episodes, and so might be able to determine whether darwinian selection or neutral drift is primarily responsible for the differences between homologous proteins. □

Methods

The protein-coding region of the primate lysozyme gene was sequenced as follows. Total cellular RNA was extracted by the single-step method²⁷ from EDTA-anticoagulated blood or frozen tissue. Most lysozymes were sequenced directly from cDNA (prepared with the GeneAmp RNA PCR kit, Perkin Elmer Cetus); this was possible because primates have one chicken-type lysozyme gene that is expressed in both the stomach and white blood cells⁵. When an RNA source was not available, the protein-coding regions were sequenced from 'long PCR' products (GeneAmp XL PCR kit, Perkin Elmer Cetus) using genomic DNA as template. DNAs were extracted from blood or frozen tissue²⁸; purified DNAs from Allen's monkey, patas, talapoin and douc langur were obtained from O. Ryder. Amplification primers were as in ref. 5, and sequencing primers were designed as needed. Sequencing was accomplished with the Taq DyeDeoxy Terminator Cycle sequencing kit and analysed on a 373A DNA sequencer (Applied Biosystems). Both strands were sequenced several times to verify the sequences carefully. If there was evidence of an allelic difference within an individual, the PCR products were cloned (TA cloning kit, Invitrogen) when necessary to obtain unambiguous sequences. The species included in this study are listed by common name, scientific name, GenBank accession number, sources (see below) and number of individuals sampled (in parentheses): human, *Homo sapiens*, J03801; chimpanzee, *Pan troglodytes*, U76912, Yerkes (1); bonobo, *Pan paniscus*, U76930-933, Yerkes (1); gorilla, *Gorilla gorilla*, U76913, Yerkes (2); orangutan, *Pongo pygmaeus*, U76914, Yerkes (1); gibbon, *Hylobates lar*, U76915, Yerkes (1); guereza colobus, *Colobus guereza*, U76916, HZG (2) and SL (1); Angolan colobus, *Colobus angolensis*, U76934-37, MMZ (2); Hanuman langur, *Semnopithecus entellus*, X60235, UCB (3); purple-faced langur, *Semnopithecus vetulus*, U76938-40, BR (1); dusky langur, *Trachypithecus obscurus*, U76917, HZG (3); Francois' langur, *Trachypithecus francoisi*, U76918, HZG (3); douc langur, *Pygathrix nemaeus*, U76941-44, SD (1); proboscis monkey, *Nasalis larvatus*, U76945-48, NY/SD (1); olive baboon, *Papio cynocephalus*, U76919, SFB (1); sooty mangabey, *Cercocebus atys*, U76920, Yerkes (1); rhesus macaque, *Macaca mulatta*, X60236, NE (1); Allen's monkey, *Allenopithecus nigroviridis*, U76949-51, SD (1); talapoin, *Miopithecus talapoin*, U76952-55, SD (1); patas, *Erythrocebus patas*, U76956-59, SD (1); vervet, *Cercopithecus aethiops*, X60237, NE (1); squirrel monkey, *Saimiri sciureus*, U76921, NE (1); cotton-top tamarin, *Saguinus oedipus*, U76922, NE (1); common marmoset, *Callithrix jacchus*, U76923, NE (1). Intron 3 of the lysozyme gene was amplified and sequenced (primer sequences are available from C.-B.S. on request) from orangutan (U76924),

Allen's monkey (U76951), mangabey (U76928), guereza colobus (U76925), and from Hanuman (U76929), dusky (U76926), Francois' (U76927) and purple-faced (U76940) langurs. The new lysozyme sequences have been deposited into GenBank under the accession numbers listed above. Institutional abbreviations are as follows, with contact persons in parentheses: NE, New England Regional Primate Research Center (D. Lee-Parritz); Yerkes, Yerkes Regional Primate Research Center (H. McClure); SD, The Zoological Society of San Diego (O. Ryder); HZG, Houston Zoological Gardens (B. Lester and J. Flanagan); UCB, University of California at Berkeley (P. Dolhinow); SL, St. Louis Zoological Park (I. Porton); SFBP, Southwest Foundation for Biomedical Research (K. Rice); MMZ, Miami MetroZoo (W. Zeigler); BR, Greater Baton Rouge Zoo (C. Lehn); NY, New York Zoological Society (G. Amato). The previously published⁵ non-human primate sequences were verified in the present study; as suggested⁵, the baboon lysozyme sequence differs slightly from the published protein sequence used in earlier analyses^{4,6}. Only in gorilla and orangutan did we find allelic differences in the coding region (G. Maston and C.-B.S., manuscript in preparation); the less-derived alleles are considered here. For each species, we sequenced the entire coding region; only the mature region is used in the present analyses. K_A and K_S values for all pairwise comparisons were calculated with the computer program Li93 (ref. 10). Similar overall results were found using the method of Nei and Gojobori^{29,30}; the results from the Li method^{9,10} are presented because the primate lysozyme sequences best meet its assumptions. Statistical significances for the K_A and K_S values were calculated by t -tests^{16,30}. For a matrix of K_A -to- K_S values, see Supplementary Information. Rate comparisons between the introns and synonymous sites were done using MEGA³⁰, and included the published human sequence.

Received 10 June; accepted 5 November 1996.

- Kreitman, M. & Akashi, H. *Annu. Rev. Ecol. Syst.* **26**, 403–422 (1995).
- Li, W.-H. & Graur, D. *Fundamentals of Molecular Evolution* (Sinauer, Sunderland, MA, 1991).
- Gillespie, J. *The Causes of Molecular Evolution* (Oxford Univ. Press, New York, 1991).
- Stewart, C.-B., Schilling, J. W. & Wilson, A. C. *Nature* **330**, 401–404 (1987).
- Swanson, K. W., Irwin, D. M. & Wilson, A. C. *J. Mol. Evol.* **33**, 418–425 (1991).
- Stewart, C.-B. & Wilson, A. C. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 891–899 (1987).
- Chivers, D. J. & Hladik, C. M. *J. Morphol.* **166**, 337–386 (1980).

- Dobson, D. E., Prager, E. M. & Wilson, A. C. *J. Biol. Chem.* **259**, 11607–11616 (1984).
- Li, W.-H., Wu, C.-I. & Luo, C.-C. *Mol. Biol. Evol.* **2**, 150–174 (1985).
- Li, W.-H. *J. Mol. Evol.* **36**, 96–99 (1993).
- Nei, M. *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).
- Endo, T., Ikeya, K. & Gojobori, T. *Mol. Biol. Evol.* **13**, 685–690 (1996).
- Wolfe, K. H. & Sharp, P. M. *J. Mol. Evol.* **37**, 441–456 (1993).
- Lee, Y.-H. & Vacquier, V. D. *Biol. Bull.* **182**, 97–104 (1992).
- Swanson, W. J. & Vacquier, V. D. *Proc. Natl Acad. Sci. USA* **92**, 4957–4961 (1995).
- Hughes, A. L. & Nei, M. *Nature* **335**, 167–170 (1988).
- Nei, M. & Jin, L. *Mol. Biol. Evol.* **6**, 290–300 (1989).
- Delson, E. in *Colobine Monkeys: their Ecology, Behaviour and Evolution* (eds Davies, A. G. & Oates, J. F.) 11–43 (Cambridge Univ. Press, 1994).
- Felsenstein, J. *PHYLP (Phylogeny Inference Package), Version 4.0* (Department of Genetics, University of Washington, Seattle, WA, 1996).
- Swofford, D. *PAUP* (Phylogenetic Analysis Using Parsimony and Other Methods)* (Sinauer, Sunderland, MA, in the press).
- Yang, Z., Kumar, S. & Nei, M. *Genetics* **141**, 1641–1650 (1995).
- Maddison, W. P. & Maddison, D. R. *MacClade: Analysis of Phylogeny and Character Evolution, Version 3.04* (Sinauer, Sunderland, MA, 1994).
- Stewart, C.-B. *Lysozyme Evolution in Old World Monkeys* (University of California, Berkeley, 1986).
- Jolles, J. et al. *J. Mol. Evol.* **28**, 528–535 (1989).
- Irwin, D. M. & Wilson, A. C. *J. Biol. Chem.* **265**, 4944–4952 (1990).
- Yu, M. & Irwin, D. M. *Mol. Phylogenet. Evol.* **5**, 298–308 (1996).
- Chomczynski, P. & Sacchi, N. *Anal. Biochem.* **162**, 156–159 (1987).
- Ausubel, F. M. et al. *Current Protocols in Molecular Biology* (Wiley, New York, 1989).
- Nei, M. & Gojobori, T. *Mol. Biol. Evol.* **3**, 418–426 (1986).
- Kumar, S., Tamura, K. & Nei, M. *MEGA: Molecular Evolutionary Genetics Analysis* (Pennsylvania State Univ., 1993).

SUPPLEMENTARY INFORMATION is available on Nature's World-Wide Web site (<http://w.w.w.nature.com>) or as paper copy from Mary Sheehan at the London editorial office of Nature.

ACKNOWLEDGEMENTS. We thank the people and institutions listed in the Methods section for providing the primate specimens used in this study; G. Maston, S.-H. Li, R. Collura, M. Auerbach and P. Banerjee for sharing unpublished sequences; and many people for providing computer programs, including D. Swofford for a test version of PAUP* 4.0, J. Felsenstein for a test version of PHYLP 4.0, Z. Yang for PAML, W.-H. Li for Li93, M. Nei for SEND and MEGA, Y. Ima for new1, and S. Muse for CODRATES. This work was supported by a grant from the NIH. C.-B.S. is an NSF Presidential Faculty Fellow.

CORRESPONDENCE and requests for materials should be addressed to C.-B.S. (e-mail: c.stewart@albany.edu). Primer sequences of intron 3 of the lysozyme gene are available from C.-B.S., and the GenBank accession numbers for the new lysozyme sequences are listed in the Methods section.

Abnormal temporal dynamics of visual attention in spatial neglect patients

Masud Husain*, Kimron Shapiro†, Jesse Martin† & Christopher Kennard*

* Department of Clinical Neuroscience, Charing Cross and Westminster Medical School, Charing Cross Hospital, London W6 8RF, UK
† School of Psychology, University of Wales, Bangor LL57 2DG, UK

WHEN we identify a visual object such as a word or letter, our ability to detect a second object is impaired if it appears within 400 ms of the first^{1–5}. This phenomenon has been termed the attentional blink or dwell time and is a measure of our ability to allocate attention over time (temporal attention). Patients with unilateral visual neglect are unaware of people or objects contralateral to their lesion^{6,7}. They are considered to have a disorder of attending to a particular location in space (spatial attention)^{6–11}. Here we examined the non-spatial temporal dynamics of attention in patients, using a protocol for assessing the attentional blink. Neglect patients with right parietal, frontal or basal ganglia strokes had an abnormally severe and protracted attentional blink. When they identified a letter, their awareness of a subsequent letter was significantly diminished for a length of time that was three times as long as for individuals without neglect. Our results demonstrate for the first time that visual neglect is a disorder of directing attention in time, as well as space.

Visual neglect is a common disorder following stroke. It is most severe after right hemisphere lesions, affecting over 70% of such patients¹². These individuals are unaware of people or objects to their left and have a poor prognosis for recovery of independent

function^{12,13}. Despite the intense interest in neglect, the mechanisms underlying this disorder remain obscure^{6,7}. One prominent theory considers left-sided neglect to be the result of a bias to attend to the right⁸. An alternative hypothesis is that neglect patients have a direction-specific impairment of disengaging attention from a stimulus on the right when they are required to shift attention to the left^{9–11}. Both theories make the same fundamental assumption that neglect is an impairment of spatial attention. We investigated whether there is a non-spatial component of attention in neglect by measuring the temporal dynamics of attention at one location. The protocol we used allowed us to measure directly the time required to discriminate an object and release processing capacity for another, when no directional shift of attention is required.

Individuals without any neurological abnormality experience a significant 'loss' of attention for ~400 ms after engaging a target for purposes of identification^{1–5}: this has been referred to as the attentional blink or dwell time. A standard procedure for determining this loss of temporal attention requires individuals to view a rapid serial visual presentation (RSVP) sequence of letters presented successively at the same location (see Fig. 1 and Methods for further details). In each RSVP sequence, all the letters are black except one, which is white. This is the first target (T1) the subject is asked to identify. In half the trials, T1 is followed at some point in the sequence by a black 'X'. This is the second target (T2). Individuals' ability to detect T2 correctly after successfully identifying T1 in this dual-target task is then plotted over time (interval between T1 and T2).

The performance of ten right-handed volunteers who had not suffered a stroke (mean age, 73 years) is shown in Fig. 2. When asked to identify T1 and also say whether T2 was present (dual-target task), their ability to detect T2 varied according to its temporal position in the RSVP sequence: if it occurred within 360 ms of T1, detection was significantly impaired. This is the standard attentional blink outcome^{1–5}. These subjects were also tested on a control task in which they viewed similar RSVP

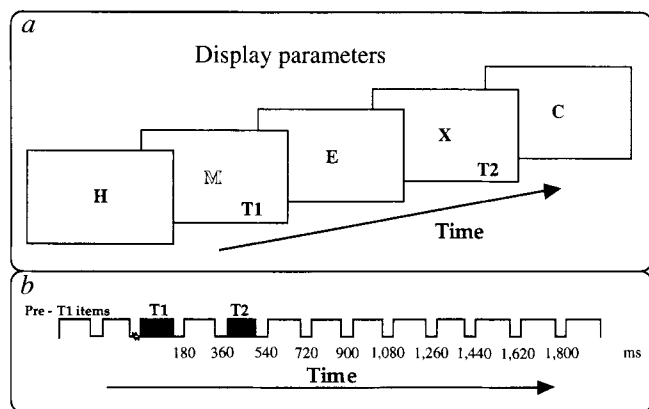


FIG. 1 *a*, Some of the stimuli in a RSVP stream. *b*, The temporal sequence of a typical RSVP trial. All sequences contained one white letter (T1), which in this case is 'M'. The number of letters presented before T1 varied randomly between 7 and 15. T2 was always a black 'X' that was present in only half the trials. In this example, it is shown as the second letter in the (10-letter) series that follows T1. In dual-target trials, subjects had to identify T1 and also say whether T2 was present. In single-target trials they had only to report the presence or absence of T2.

sequences but simply had to detect the presence of T2, ignoring T1. In this single-target task, they detected T2 correctly on >90% of trials, regardless of when it occurred in the sequence after T1 (Fig. 2). This shows that the attentional blink (observed in the dual-target task) is not due simply to a failure to maintain attention throughout the duration of the RSVP sequence.

When eight right-handed right-hemisphere stroke patients without neglect (mean age, 64 years; mean lesion volume, 23.5 cm³) were tested (mean, 26 days after stroke), their attentional blink also lasted only 360 ms and was not significantly different in magnitude from that of the volunteers without stroke (Fig. 2). Neither was their performance different from normal volunteers on the single-target task during the interval associated with the attentional blink. Lesions defined by computed tomography were plotted using the templates of Damasio and Danasio¹⁴. Four of these patients had suffered infarction of

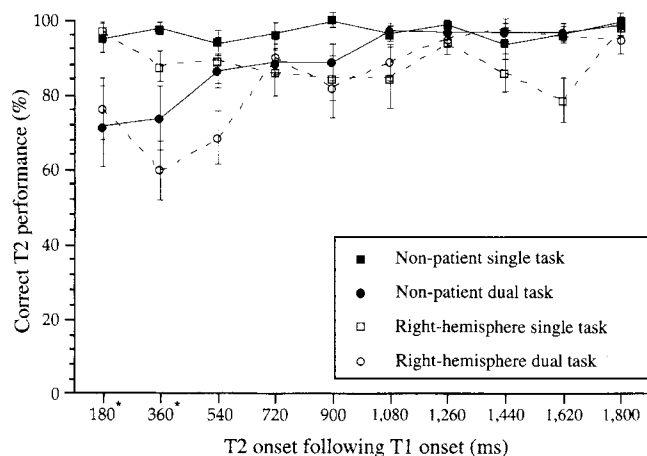


FIG. 2 Performance of normal individuals without stroke (filled symbols) and right-hemisphere stroke patients without neglect (open symbols). In both groups, T2 detection on the dual-target task (circles) was significantly different from that on the single-target (squares) task for T1–T2 intervals up to 360 ms (indicated by asterisked times on the abscissa; $P < 0.05$). Error bars represent standard errors of the mean. T1 was correctly identified on 86 and 81% of dual-target trials, respectively, by normal volunteers and those with stroke.

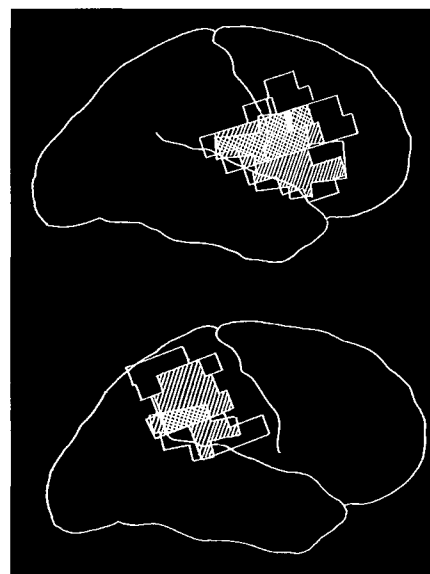


FIG. 3 Extent of cortical lesions in seven patients with left-sided visual neglect. Single hatching represents the region of overlap of 2 lesions; cross-hatching shows the overlap of 3 lesions; solid white is the zone of overlap of 4 lesions. The two-lesion foci are located in the inferior parietal lobe and in the inferior frontal lobe. One patient suffered a haemorrhage of the basal ganglia without cortical involvement (lesion not shown).

cortical and subcortical regions: two of the superior parietal lobe, one of the temporal lobe, and one of the medial frontal lobe. The other four had suffered subcortical strokes, but none had only a pure lacunar syndrome.

The performance of eight right-handed patients (mean age, 64 years) with left-sided visual neglect following right hemisphere stroke (mean lesion volume, 37 cm³, not significantly different from patients without neglect; $t = 1.6$, $P > 0.05$) was then assessed (mean, 34 days after stroke, not significantly different from patients without neglect; $t = 0.6$, $P > 0.05$). Three subjects had suffered parietal stroke, with the region of greatest lesion overlap being confined to the inferior parietal lobe, the cortical area most commonly associated with neglect^{15,16} (Fig. 3). Four patients had frontal infarcts, with the region of greatest lesion overlap in the inferior frontal lobe (Fig. 3). This region (part of the homologue of Broca's area in the left hemisphere) has been found to be the critical lateral frontal area associated with neglect^{17,18}. One subject had suffered a haemorrhage of the basal ganglia.

All patients had visual neglect clinically and on the Mesulam shape-cancellation task¹⁹. The neglect patients found a mean of only 24 targets (range, 6–36) out of 60, all on the right side of the sheet. By comparison, the patients without neglect had found a mean of 57 out of 60 targets (range, 54–60), without any spatial bias in their omissions.

When patients with neglect were required to identify T1 and detect T2 (dual-target task), they were severely impaired in their ability to detect T2. Performance was worst (42% correct) when T2 appeared 180 ms after T1 and improved progressively as the interval between T1 and T2 increased, until it returned to baseline (single-target task) performance by 1,440 ms (Fig. 4). The attentional blink in visual neglect is thus of significantly greater magnitude and significantly more protracted than in normal subjects or stroke patients without neglect. Performance did not differ significantly between patients with parietal, frontal or basal ganglia lesions. In the single-target task, neglect patients detected T2 correctly on 80–95% of trials, regardless of when it was presented in the RSVP sequence. This did not differ significantly from the performance of patients without neglect. Overall, there was a significant correlation between Mesulam cancellation score and magnitude of attentional blink (integral or area between

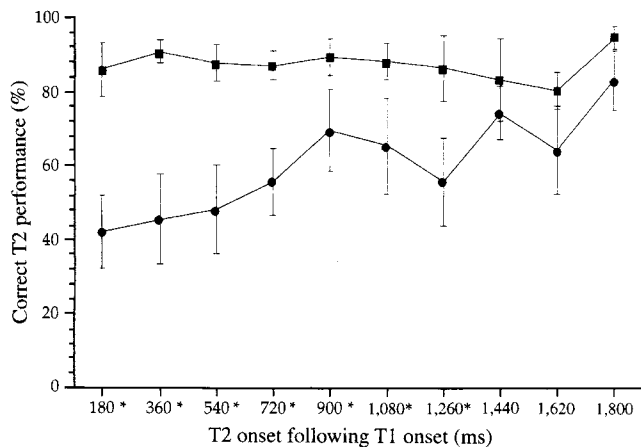


FIG. 4 Performance of patients with visual neglect. Detection of T2 on the dual-target task (circles) was significantly different from that on the single-target task (squares) for T1–T2 intervals less than 1,440 ms (indicated by asterisks on the abscissa; $P < 0.05$). Furthermore, the magnitude of visual unawareness during this attentional blink was much greater than in stroke patients without neglect or in normal individuals. T1 was correctly identified on 72% of dual-target trials.

dual- and single-task performance curves; $r = -0.71$, $n = 16$, $P < 0.05$).

What causes the temporal impairment in neglect? Studies of the attentional blink in normal individuals have raised the possibility of at least two types of underlying mechanism^{1–5}. First, the visual system may take time correctly to conjoin or bind features of the target (that is, white colour with the shape of the letter to which it belongs) before features belonging to the subsequent item can be processed (compare with refs 20 and 21). Alternatively, there may be competition between critical stream items for selection after features belonging to each item have been conjoined correctly (compare with ref. 22). So far, the evidence in normal individuals suggests that the latter is a more likely explanation^{3,5}. By contrast, evidence for an impairment of feature binding in an individual with bilateral parietal lesions has been presented²³. Furthermore, positron emission tomography has demonstrated increased activity in parietal cortex when conjunctions between features are required²⁴. There is little evidence to suggest a role for the frontal lobe in feature binding, but abnormal temporal dynamics of attention have been reported in subjects following frontal lobe surgery²⁵.

Whatever the exact mechanism responsible for the abnormal

attentional blink, the results presented here require a reformulation of prevailing spatial accounts of visual neglect. Previous studies have demonstrated impairments of spatial attention in this condition^{9–11,26}. We have shown that once attention is committed to the analysis of a visual object, there is an impairment in the ability to direct it to another, even if both stimuli are presented at the same location. We suggest neglect has two components. First, there is a spatial bias to direct attention towards stimuli processed by the undamaged cerebral hemisphere, and second, there is a deficit in temporal processing, regardless of where attention is directed. Our finding may also help to explain a non-spatial disorder of visual attention associated with bilateral lesions of the posterior parietal lobe—simultagnosia in Bálint's syndrome^{27,28}. Individuals with this condition are able to attend to only one of two overlapping objects presented simultaneously at the same location, even though they have no apparent deficit when attending to either of these objects presented alone²⁹. In this case, we suggest that there is no spatial bias because both hemispheres are damaged. Nevertheless, a fundamental disorder of visual processing remains: patients experience profound difficulty engaging a second stimulus after engaging a first. We suggest that this is a severe form of the temporal component of neglect. □

Methods

All subjects viewed from a distance of 35 cm a rapid serial visual presentation (RSVP) stream of letters presented at the centre of a computer monitor. Each letter was presented for 131 ms, with an interstimulus interval of 49 ms, yielding a presentation rate of 5.6 letters per second (Fig. 1). Letters subtended $\sim 1.4^\circ$ of visual angle in height and 0.7° in width. All letters were black except the first target letter (T1), which was white. The background was a uniform grey field, present throughout the sequence. Subjects initiated a trial when ready by depressing a computer mouse button. Each trial began with a 360-ms presentation of a small white fixation dot. The number of letters presented before T1 varied randomly between 7 and 15. T1 appeared on all trials and was always followed by a sequence of ten letters (Fig. 1). T1 could be any letter in the alphabet except 'X'. The second target (T2) was a black 'X'. It appeared on a randomly chosen half of the trials, and was presented randomly as one of the ten letters that followed T1. An 'X' was never presented before T1 and never appeared twice within a single RSVP stream.

Subjects were instructed in separate trial blocks to report the presence or absence of T2 alone (single-target trials), or to identify T1 and then also report on whether T2 was present (dual-target trials). T2 was presented eight times at each of the 10 possible serial positions after T1, yielding a total of 80 T2-present trials and 80 T2-absent trials for each condition (single or dual target). Subjects received 20 practice trials in each condition before data collection. The order of conditions was counterbalanced across subjects, and stroke patients performed either the single- or dual-target task on two different days. The interval between the two tests was less than a week in all cases. In each condition, subjects performed eight blocks, each consisting of 20 trials (10 with T2 present and 10 with T2 absent). Within a block, T2 was randomly presented once at each of the ten serial positions following T1. Short rest breaks were taken between each block.

Received 28 August; accepted 5 November 1996.

- Broadbent, D. E. & Broadbent, M. H. P. *Percept. Psychophys.* **42**, 105–113 (1987).
- Raymond, J. E., Shapiro, K. L. & Arnell, K. M. *J. exp. Psychol., hum. Percept. Perform.* **18**, 849–860 (1992).
- Shapiro, K. L., Raymond, J. E. & Arnell, K. M. *J. exp. Psychol., hum. Percept. Perform.* **20**, 357–371 (1994).
- Duncan, J., Ward, R. & Shapiro, K. *Nature* **369**, 313–315 (1994).
- Shapiro, K. L. & Raymond, J. E. in *Inhibitory Processes in Attention, Memory, and Language* (eds Dagenbach, D. & Carr, T.) 151–188 (Academic, New York, 1994).
- Robertson, E. H. & Marshall, J. C. (eds) *Unilateral Neglect: Clinical and Experimental Studies* (Erlbaum, Hillsdale, NJ, 1993).
- Jeannerod, M. (ed.) *Neurophysiological and Neuropsychological Aspects of Spatial Neglect* (North-Holland, Amsterdam, 1987).
- Kinsbourne, M. *Trans. Am. Neurol. Assoc.* **95**, 143–145 (1970).
- Posner, M. I., Walker, J. A., Friedrich, F. J. & Rafal, R. D. *J. Neurosci.* **4**, 1863–1874 (1984).
- Posner, M. I., Walker, J. A., Friedrich, F. A. & Rafal, R. D. *Neuropsychologia* **25**, 135–145 (1987).
- Morrow, L. A. & Ratcliff, G. *Psychobiology* **16**, 261–269 (1988).
- Stone, S. P., Patel, P. & Greenwood, R. J. *J. Neurol. Neurosurg. Psychiatr.* **56**, 463–466 (1993).
- Fullerton, K. J., Mackenzie, G. & Stout, R. W. Q. *J. Med.* **250**, 147–162 (1988).
- Damasio, H. & Damasio, A. R. *Lesion Analysis in Neuropsychology* (Oxford University Press, New York, 1989).

- Vallar, G. & Perani, D. *Neuropsychologia* **24**, 609–622 (1986).
- Heilman, K. M., Watson, R. T. & Valenstein, E. in *Localization and Neuroimaging in Neuropsychology* (ed. Kertesz, A.) 495–524 (Academic, San Diego, 1994).
- Husain, M. & Kennard, C. J. *Neurol.* **243**, 652–657 (1996).
- Jackson, S. & Husain, M. *Curr. Opin. Neurobiol.* (in the press).
- Mesulam, M.-M. *Principles of Behavioural Neurology. Tests of Directed Attention and Memory* (Davis, Philadelphia, 1985).
- Treisman, A. & Gelade, G. *Cogn. Psychol.* **12**, 97–136 (1980).
- Treisman, A. *Curr. Opin. Neurobiol.* **6**, 171–178 (1996).
- Duncan, J. & Humphreys, G. W. *Psychol. Rev.* **96**, 433–458 (1989).
- Friedman-Hill, S. R., Robertson, L. C. & Treisman, A. *Science* **269**, 853–855 (1995).
- Corbetta, M., Shulman, G. L., Miezin, F. M. & Petersen, S. E. *Science* **270**, 802–805 (1995).
- Richer, F. & Lepage, M. *Neuropsychologia* **34**, 509–514 (1996).
- Riddoch, M. J. & Humphreys, G. W. *Neuropsychologia* **21**, 589–599 (1983).
- Bálint, R. *Orvosi Hetilap*, 209–236 (1907).
- Husain, M. & Stein, J. *Arch. Neurol.* **45**, 89–93 (1988).
- Humphreys, G. W., Romani, C., Olson, A., Riccochi, M. J. & Duncan, J. *Nature* **372**, 357–359 (1994).

ACKNOWLEDGEMENTS. M.H. is supported by a grant from the Wellcome Trust. We thank J. Raymond, S. Tipper, S. Treue and R. Ward for comments on the manuscript.

CORRESPONDENCE and requests for materials to M.H. (e-mail: m.husain@cxwms.ac.uk).

Visuomotor integration is associated with zero time-lag synchronization among cortical areas

Pieter R. Roelfsema^{*†}, Andreas K. Engel^{*}, Peter König[‡] & Wolf Singer^{*}

^{*} Max-Planck-Institute for Brain Research, Deutschordenstrasse 46, 60528 Frankfurt, Germany

[†] The Netherlands Ophthalmic Research Institute, and University of Amsterdam, Department of Medical Physics, P.O. Box 12141, 1100 AC Amsterdam, The Netherlands

[‡] The Neurosciences Institute, 10640 John Jay Hopkins Drive, San Diego, California 92121, USA

INFORMATION processing in the cerebral cortex invariably involves the activation of millions of neurons that are widely distributed over its various areas. These distributed activity patterns need to be integrated into coherent representational states. A candidate mechanism for the integration and coordination of neuronal activity between different brain regions is synchronization on a fine temporal scale¹⁻³. In the visual cortex, synchronization occurs selectively between the responses of neurons that represent related features²⁻⁵ and that need to be integrated for the generation of coherent percepts; neurons in other areas of the cerebral cortex also synchronize their discharges⁶⁻¹⁰. However, little is known about the patterns and the behavioural correlates of synchrony among widely separated cortical regions. Here we report that synchronization occurs between areas of the visual and parietal cortex, and between areas of the parietal and motor cortex, in the awake cat. When cats responded to a sudden change of a visual pattern, neuronal activity in cortical areas exhibited synchrony without time lags; this synchrony was particularly strong between areas subserving related functions. During reward and inter-trial episodes, zero-time-lag synchrony was lost and replaced by interactions exhibiting large and unsystematic time lags.

The visual cortex is composed of a large number of areas^{11,12}, which contain neurons that are tuned to different visual features and that have different roles in the analysis of visual information¹³. Similar organizational principles hold for the parietal and the motor cortex, which also consist of numerous areas¹¹⁻¹⁶. Any cerebral activity involves large numbers of areas, which must

coordinate their activity. This coordination can be investigated by studying the correlations between field potentials, which reflect the average activity of large groups of neurons in the vicinity of a recording electrode¹⁷. The strength of coupling between transcortically recorded field potentials in different cortical areas changes dynamically during the performance of a behavioural task¹⁰. However, the spatial pattern of interactions between different areas and the time lags in these interactions have not yet been evaluated systematically.

Five cats were trained to press and release a lever with their forepaw in response to visual stimuli. At the onset of a behavioural trial, a grating was presented within a circular aperture. When the orientation of the grating changed, the cats had to respond by either pressing or releasing the lever, and were subsequently rewarded with food. When the cats performed this task reliably, transcortical electrodes were implanted under general anaesthesia. Coupling among transcortical field potentials that were recorded during the various episodes of the behavioural task was investigated using cross-correlation analysis.

Interactions between areas changed in a characteristic and highly reproducible manner during the course of the task. This is exemplified in Fig. 1 for the interaction between two areas of the parietal association cortex, area 7 and the lateral subdivision of area 5 (area 5l). During the entire task period, field potentials at the two recording sites were synchronized precisely, as indicated by a pronounced peak in the correlation function at zero time lag (Fig. 1a-c). The strength of this correlation varied between 24% and 33%, and was highest shortly after the cat pressed the response lever (Fig. 1b, e). The spacing of 46 ms between subsequent peaks of the correlation functions indicates that coupling was particularly strong for frequencies in the mid-beta frequency range (20-25 Hz), which dominate the power spectrum of the field potentials (Fig. 1f). During the reward period, large time lags of -7 to +50 ms occurred in the correlation function (Fig. 1d, e). This loss of zero-time-lag synchrony was associated with a strong increase of power in the alpha-band (7-13 Hz) (Fig. 1g).

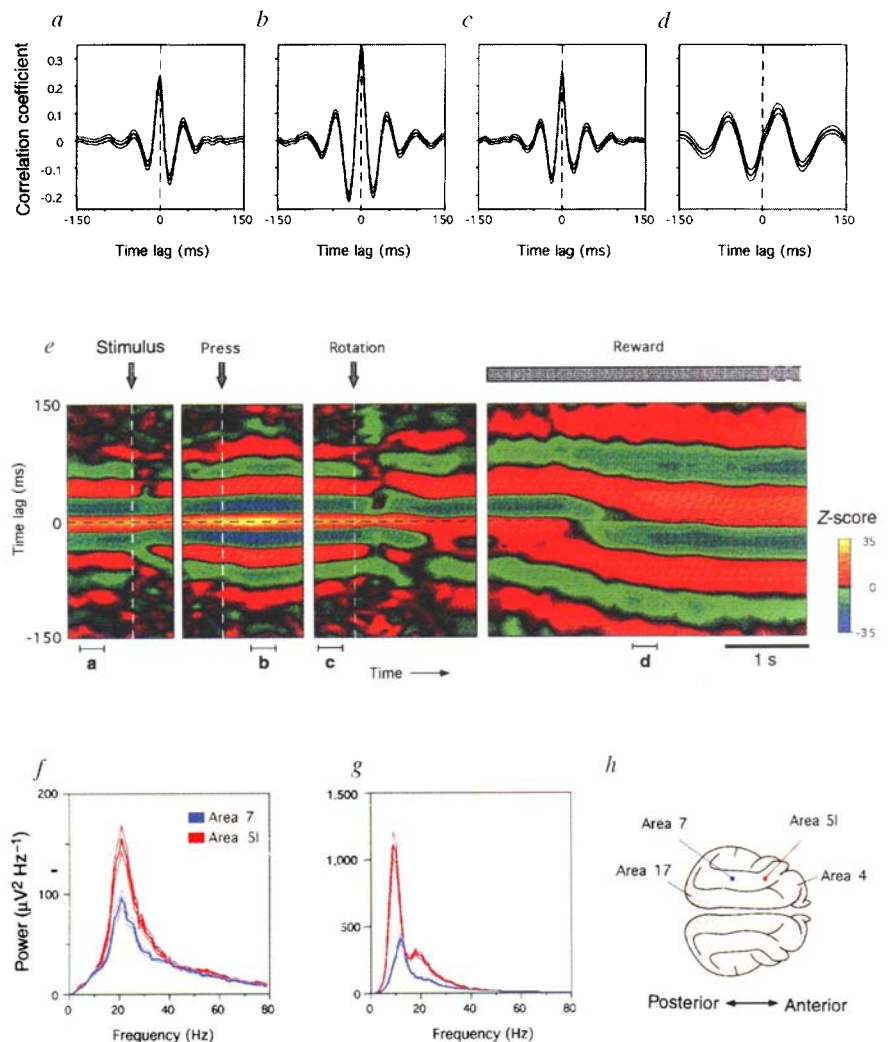
A predominance of alpha-rhythms was a characteristic feature of the field potentials in all cats during the reward period (Fig. 2d). The amount of power in the alpha-band was on average more than fivefold greater during the reward period than when the animals paid attention to the visual stimuli ($P < 0.001$, *U*-test). Time lags in the correlation functions were compared between stimulus and reward periods for all area combinations exhibiting a robust interaction (*Z*-score > 6). On average, time lags in the interactions between areas were significantly smaller when the cats performed the task than during the reward period (Fig. 2d, $P < 0.00001$, *U*-test). When the cats were in the test box, but not involved in their task, low-frequency field-potential rhythms

TABLE 1 Strength of zero-time-lag synchronization between cortical areas

Correlation coefficient				Correlation coefficient			
Areas	(%)	Range	N	Areas	(%)	Range	N
17c-7c	9 ± 4	4-14	4	5lc-4c	2 ± 3	0-5	2
17c-5lc	0		2	5lc-17i	0		2
17c-5mc	0		3	5lc-18i	4 ± 5	0-8	2
17c-4c	0		2	5lc-21i	4 ± 4	0-8	3
17c-17i	22 ± 7	12-28	4	5mc-4c	10 ± 1	9-11	2
17c-18i	12 ± 3	7-14	4	5mc-17i	0		3
17c-21i	8 ± 2	7-10	3	5mc-18i	0		3
7c-5lc	17 ± 10	10-24	2	5mc-21i	2 ± 3	0-6	4
7c-5mc	7 ± 2	6-9	3	4c-17i	0		2
7c-4c	0		2	4c-18i	0		2
7c-17i	6 ± 2	5-8	4	4c-21i	0		2
7c-18i	9 ± 5	5-15	4	17i-18i	20 ± 3	17-24	4
7c-21i	12 ± 2	9-14	3	17i-21i	21 ± 10	12-32	3
5lc-5mc	14 ± 6	9-21	3	18i-21i	15 ± 2	14-18	3

Correlation coefficients (±s.d.) are averaged across cats. Strength of the zero-time-lag synchronization for the epochs when the cats waited for rotation of the grating. Abbreviations: i, areas ipsilateral to the forepaw used by the cats; c, areas contralateral to the forepaw used by the cats; N, number of cats.

FIG. 1 Changes in the interactions between cortical areas during task performance. *a–d*, Correlation functions between field potentials recorded in parietal area 7 and the lateral subdivision of area 5, averaged over trials. Black curves represent correlation functions that were computed for the following periods: *a*, when the cat waited for the appearance of the visual stimulus; *b*, an epoch shortly after the animal pressed the response lever; *c*, the phase in which the animal waited for the rotation of the stimulus; and *d*, the reward epoch. Curves on either side demarcate the 95% confidence interval of the correlation functions (± 2 s.e.m.). Positive time lags in the correlograms indicate that activity in area 5I leads over activity in area 7. The ordinate shows correlation coefficients. *e*, Alterations in the average correlation function during the task. Abscissa, time relative to the behavioural event on which computational windows were aligned; ordinate, time lag of the correlation function. Computational windows were aligned on the onset of the visual stimulus in the first panel; on the pressing of the response key in the second panel; on the rotation of the grating in the third panel; and on the delivery of a food reward in the fourth panel. The computational windows corresponding to the correlation functions in (*a–d*) have been indicated below the respective panels. The colour scale indicates the Z-score of the correlation (value of the average correlation function normalized to the s.e.m.): red–yellow, positive correlations; green–blue, negative correlations. *f*, Power spectrum of the field potential recorded from area 7 (blue curve) and area 5I (red curve) in a 1-s window when the animal was pressing the response key. Lighter curves show the respective 95% confidence intervals. *g*, Power spectra at the two recording sites during the reward epoch (note the different scale on the ordinate). *h*, Location of the two recording sites in the left hemisphere. The location of the primary visual cortex (area 17) and the primary motor cortex (area 4) have been shaded.



occurred spontaneously and were also associated with large time lags in the correlation functions. This indicates that synchrony among cortical areas is tighter when animals are engaged in a visumotor task that requires attention and prompt reaction than when they are engaged in feeding or are at rest.

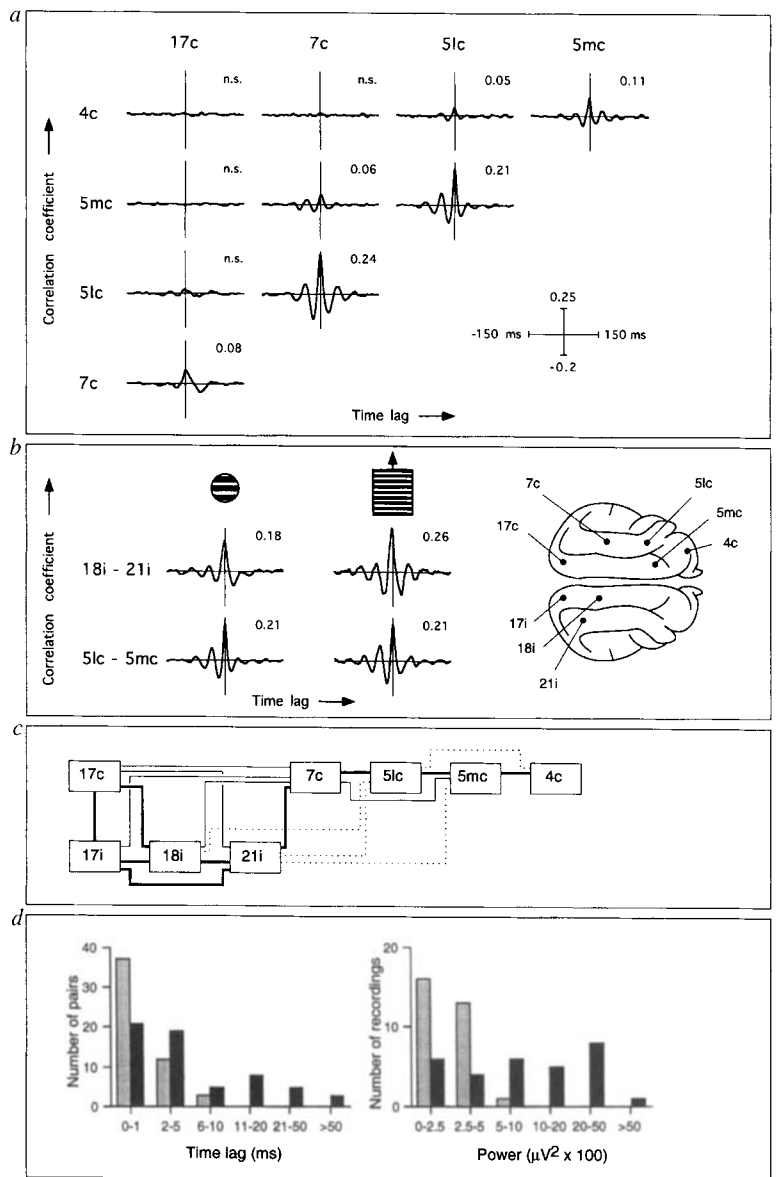
Correlations among field potentials recorded from areas of the visual, parietal and motor cortex when the cat was paying attention to the grating are shown in Fig. 2a. All significant interactions between areas consisted of synchrony with a time lag of approximately 0 ms (Fig. 2a, b, d). Significant synchronization occurred between area 17, the primary visual cortex, and area 7 of the parietal cortex. Area 7, in turn, exhibited significant synchronization with the lateral subdivision of area 5 and slightly weaker synchronization with the medial subdivision of area 5 of the parietal cortex. Conversely, area 4 of the motor cortex exhibited stronger synchronization with the medial than with the lateral subdivision of area 5. Thus synchronization did not occur between areas that are functionally remote (Fig. 2a). During the task period, the pattern of synchronization among visual areas was strongly dependent on the type of visual stimulation (Fig. 2b). When the visual stimulus was changed from a stationary to a moving grating, the strength of synchrony among visual areas increased by 23% on average. The interactions among areas of the parietal and motor cortex were not altered by this change of the visual stimulus (Fig. 2b).

The average strength of zero-time-lag synchronization is listed for all combinations of investigated areas in Table 1 and shown schematically in Fig. 2c. This analysis shows that the strength of

synchrony reflects the functional relations among cortical areas rather than their spatial vicinity, thus excluding current spread from remote sources as a cause for zero-time-lag synchrony. Synchronization between parietal area 7 and contralateral area 21, an area that occupies a relatively high position in the hierarchy of visual areas^{11,12}, was stronger than synchrony between area 7 and the ipsilateral primary visual cortex (area 17), despite the latter areas having a much smaller spatial separation (Table 1 and Fig. 2c).

In many cases, multi-unit activity (MUA), action potentials generated by local clusters of neurons, could be recorded through the transcortical electrodes. In these cases, coupling between MUA and field potentials was investigated by computing spike-triggered field-potential averages (STAs), which reveal the average field-potential fluctuation around the time that an action potential was recorded. The coupling between MUA recorded in parietal area 7 and field potentials that were recorded from the same electrode and from other electrodes in parietal area 5 and visual area 17 is shown in Fig. 3. During the trial, MUA in area 7 was synchronized to a sharp field-potential fluctuation at the same electrode with a time lag close to 0 ms (Fig. 3a, b). The STA during the reward period exhibited additional peaks with a spacing of ± 100 ms, which indicates that the local cluster of neurons was entrained by the alpha-rhythm that occurred during feeding (Fig. 3c). The STAs for combinations of different areas did not reach statistical significance before the onset of the trial (Fig. 3a). Shortly after the cat pressed the response lever, however, the field potential in area 5 exhibited significant synchroni-

FIG. 2 Pattern of interactions among areas of the visual, parietal and motor cortex. *a*, Correlation functions among areas of the left hemisphere of one of the cats for the episode in which the animal was waiting for the rotation of the grating, averaged over trials. Positive time lags indicate lead of activity in the area shown on the left of the correlation functions over the area shown above the correlation functions. In all cats, recordings from somatosensory and motor areas were made in the hemisphere contralateral (c) to the paw that the animal used for pressing the lever; l, lateral; m, medial. *b*, Dependence of the correlation functions on the type of visual stimulation. Left, correlation functions obtained during visual stimulation with a stationary grating patch (18° of visual angle, $0.1 \text{ cycles deg}^{-1}$). Top row, correlation between field potentials in visual areas 18 and 21, ipsilateral (i) to the paw the cat used. Bottom row, correlations between field potentials in the lateral (5l) and medial (5m) subdivision of area 5 contralateral to the paw the cat used. Middle, correlation functions obtained during stimulation with a moving grating within a larger aperture ($31^\circ \times 42^\circ$, $0.28 \text{ cycles deg}^{-1}$, 2.5 Hz). Synchronization between area 18 and area 21 was stronger during stimulation with the moving grating than during stimulation with the stationary grating ($P < 0.0001$, *U*-test). The strength of synchrony between the two subdivisions of area 5 was not affected by this change of the visual stimulus ($P > 0.5$). Right, location of all recording sites ipsilateral and contralateral to the paw the cats used. *c*, Synchronization strength among the areas from which recordings were obtained during the task period, averaged across cats. Lines between areas indicate the strength of synchrony. Thick lines indicate synchronization with a correlation coefficient exceeding 10%, thin lines interactions with a correlation coefficient between 5% and 10%, and broken lines interactions with a correlation coefficient smaller than 5%. Areas that are not connected did not exhibit significant synchronization in any of the cats ($Z\text{-score} < 6$). Areas in the hemisphere contralateral and ipsilateral to the paw the cats used are shown in the top and bottom row, respectively. Note that the interhemispheric asymmetry in the diagram reflects an asymmetry in electrode placement, not in the pattern of synchrony. *d*, Left, distribution of time lags in the correlation functions. Grey bars show time lags when the cats waited for the rotation of the grating, black bars show time lags during the reward epoch. Right, distribution of integrated power in the alpha frequency range (7–13 Hz) when the cats waited for rotation of the grating stimulus (grey bars) and in the reward epoch (black bars).



zation to the MUA in area 7 (Fig. 3b). During feeding, neurons in area 7 fired in synchrony with the alpha-rhythms occurring in area 17 and area 5, albeit at different phases (Fig. 3c). Neurons in area 7 did not exhibit large differences in firing rates between behavioural epochs (Fig. 3d). Nonetheless, the STAs revealed large changes between behavioural conditions in interactions both within (in accordance with ref. 9) and between areas. STAs could only be computed for electrode combinations for which MUA of sufficient quality was recorded. In these cases ($N = 39$), changes in the STAs closely resembled the changes in the correlations between field potentials, demonstrating that local clusters of neurons fired in synchrony with both the local and distant field-potential rhythms.

Synchronization among neuronal responses occurs within cortical regions as diverse as the visual^{4,5}, sensorimotor⁶⁻⁸ and pre-frontal cortex⁹. It has been shown that the strength of coupling between these widely spaced cortical areas changes dynamically during task performance¹⁰. Our results are in agreement with these findings, and demonstrate that, during attentive processing, interactions between areas are characterized by tight synchronization of activity with close to zero time lag. Moreover, in this behavioural state, zero-time-lag synchrony occurs not only among areas of the visual cortex¹⁸⁻²⁰, but also between areas of the visual and parietal cortex, and between areas of the parietal and motor

cortex. It has been suggested that zero-time-lag synchronization among the distributed responses of visual cortical neurons could serve their integration into coherent representational states¹⁻³. If this is the case, the synchronization between neurons in parietal and motor cortical areas could, by analogy, serve to integrate distributed pre-movement activity into a coherent representation of a compound movement²¹.

Synchrony between areas in different hemispheres has been shown to depend on the integrity of cortico-cortical connections^{22,23}, which suggests a minor contribution from subcortical projections. The scheme of coupling between areas suggested by our results is also in agreement with the pattern of cortico-cortical connections¹¹. However, the dependence of the interactions among visual areas on the type of visual stimulation (Fig. 2b) demonstrates that the pattern of synchrony is not merely a reflection of the topology of connectivity between areas. This is in accordance with previous studies on the stimulus dependence of synchrony (reviewed in ref. 2). A similar conclusion follows from the drastic changes that occurred in the interactions during the course of a behavioural trial. Large time lags in the interactions between areas occurred only during the reward periods. The alpha-rhythms that characterize the cortical field potentials during feeding have been observed previously and were called 'post-reinforcement synchronization'²⁴, but phase relations

between field potentials in different areas during these rhythms had to our knowledge not been studied before.

The global changes in the pattern of interactions between areas may depend on the activity of modulatory projections that ascend from the brainstem, the basal forebrain and nonspecific thalamic nuclei. Activation of the mesencephalic reticular formation gives rise to behavioural arousal, increases the amount of power in the beta (15–30 Hz) and gamma (>30 Hz) frequency band of the EEG²⁵, and has been shown to enhance the occurrence of zero-time-lag synchronization of neuronal responses both within and across cortical areas^{26,27}. Our results extend these findings by showing that such shifts in the pattern of interactions do not only occur during gross, long-term changes of vigilance, like the transition from sleeping to waking. Robust changes in interactions between areas also occur in relation to variations in the attentive state of awake animals during the repetitive alternation of reward episodes and episodes in which they have to be prepared to make a swift response to a visual stimulus. □

Methods

Behavioural task. The cats were kept unrestrained in a test box, the front of which consisted of a back-projection screen. Visual stimuli were generated with a d.c.-driven overhead projector and LCD display controlled by a personal computer. The cats were separated from the screen by a transparent response key. The cats were mildly food deprived (<10% weight loss) and trained to respond to a stationary, circular patch of grating (18° of visual angle, 0.1

cycles deg⁻¹). Two cats responded to the grating stimulus by pressing the response lever, and holding the lever until the grating rotated by 90° or 180°, which occurred after 1–2.5 s. The rotation indicated that the key had to be released within 0.75 s. After correct performance the animals were rewarded with fluid cat food. For the other three cats the procedure was slightly modified to speed up learning: the onset of the grating stimulus indicated the start of a trial, but they only needed to respond to the rotation of the grating (after 2–4 s) by pressing the response key. All cats carefully attended the visual stimulus in the interval during which it was present.

Implantation of transcortical electrodes. Transcortical electrodes were implanted chronically under general anaesthesia, which was induced with ketamine and xylazine (10 and 2 mg per kg, intramuscular, respectively) and was maintained after intubation by ventilating the cat with a mixture of 70% N₂O and 30% O₂ supplemented with halothane (0.8–1.2%). Small trepanations were made and the dura was opened. Transcortical electrodes were composed of two teflon-coated platinum–iridium wires 120 µm in diameter. Insulation was removed over a distance of 80–120 µm, resulting in an impedance of 0.15–0.3 MΩ (at 100 Hz). The tip of the shorter wire was positioned in the superficial layers of the cortex, and the tip of the longer wire 1.8 mm deeper, in layer 6 or in the white matter. The electrode was glued to the pial surface with a small drop of tissue glue (cyanoacrylate), and the trepanation was closed with dental acrylic. Electrodes in area 4 were positioned in the representation of the contralateral forepaw²⁸, and electrodes in visual cortical areas according to retinotopic maps. Electrode positioning in areas 7 and 5 was guided by cytoarchitectonic maps. Electrodes were positioned at the same stereotactic coordinates in all cats. The cats were allowed to recover for at least 14 days before data collection. After completion of the recording sessions, the cats were killed with an overdose of pentobarbital and perfused transcardially. The positions of the recording sites were confirmed in Nissl-stained sections.

Correlation analysis. Cats completed on average 78 successful trials per day, and recording sessions were continued until data from at least 600 trials were obtained. Transcortical potentials were amplified differentially and filtered between 5 and 100 Hz. For computation of the correlation functions, data segments of different trials were first aligned on the events that occurred during the course of a trial (onset and rotation of the grating, time of key-press and onset of the reward epoch). Correlation functions were calculated for each trial in successive computational windows of 300 ms duration that overlapped each other by 200 ms. Subsequently, these correlation functions were averaged over trials and the variance was computed at each point of the correlation functions. Z-scores were obtained by dividing the average correlation functions by the s.e.m. Correlation functions with $Z > 6$ were considered to be significant (two-sided t -test, $P < 0.0005$ after correction for multiple comparisons (301 points per correlation function times 80 correlation functions) by the Bonferroni method). To estimate the volume of cortex that contributed to the signals recorded by a transcortical electrode, cats were anaesthetized with a mixture of 70% N₂O and 30% O₂ supplemented with halothane (0.4%), and paralysed with pancuronium bromide (0.15 mg per kg body weight per hour). Evoked potentials to a patch of square-wave grating (spatial frequency, 0.59 cycles deg⁻¹; size, $1.7^\circ \times 1.7^\circ$) that reversed phase at 10 Hz were recorded from the transcortical electrodes in area 17. At a receptive field eccentricity of 3° the amplitude of the evoked potential was reduced to less than 50% when the grating patch was shifted by 1.7° . Taking the cortical magnification factor at this eccentricity into account²⁹, this shift corresponds to a cortical distance of less than 1.5 mm. Because the minimal distance between transcortical electrodes was 5 mm, a contribution of current spread (volume conduction) to the correlation functions can be excluded. The same conclusion follows from the fact that correlation functions exhibited small time lags and asymmetrical satellite peaks during the task period and larger time lags during the reward period, even for the electrode combinations with the smallest separation. In Fig. 2b, for example, the field potential in area 18 exhibited a small (2 ms) time lead over that in area 21, and asymmetrical satellite peaks occurred in the correlation function between the lateral and medial subdivision of area 5. Because volume-conducted sources would have contributed without time lag to both recording sites that were subjected to correlation analysis, the observed time lags and asymmetrical satellite peaks exclude a volume-conducted contribution to the correlation functions.

Spike-triggered field-potential averages (STAs). Although electrodes remained *in situ* for more than 6 months, stable MUA recordings with a signal-to-noise ratio larger than 3 were obtained from some of the wires. STAs were computed for the spikes in windows of 500 ms duration in successive trials from a single recording day. On average, correlations among field potentials were more pronounced than those revealed by STAs in inter-areal combinations. Several factors probably contribute to this difference. First, changes in the trigger levels for spike detection and small drifts in the position of the wires precluded pooling of STAs over days, whereas correlations between field potentials were highly reproducible and could be pooled. Second, MUA reflects only suprathreshold activity of cells near the electrode tip, whereas field potentials result from the average sub- and supra-threshold responses of neurons within a cortical volume with radius over several hundred micrometres (ref. 17). Third, field potentials are biased towards representing activity that is locally synchronous¹, whereas MUA recordings are unbiased in this respect. It

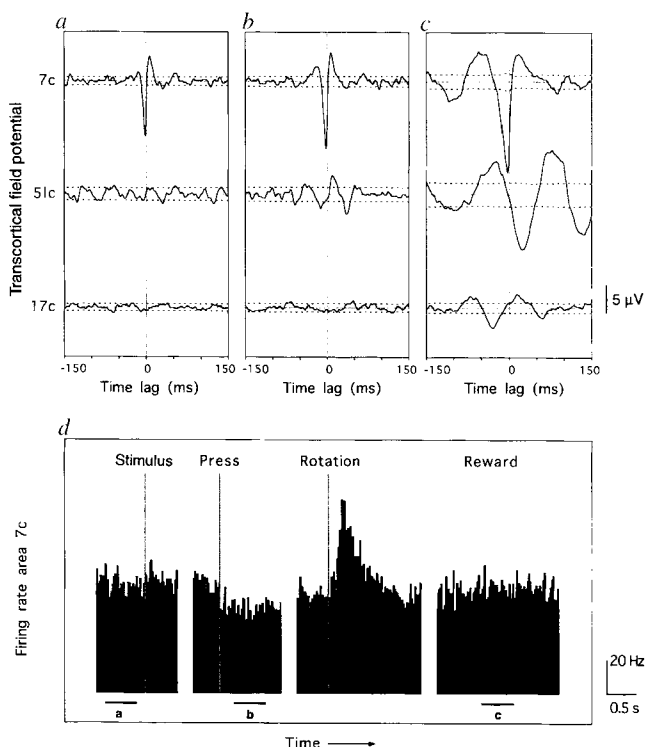


FIG. 3 Alterations in spike-triggered field-potential averages during the behaviour trial. *a–c*, Spike-triggered averages (STAs) obtained by correlating MUA in area 7 to the field potential at the same electrode (top trace), the field potential in area 5 (middle trace) and that in area 17 (bottom trace). STAs were computed: *a*, when the cat waited for the appearance of the visual stimulus; *b*, after the cat pressed the response lever; and *c*, in the reward epoch. Depth positivity (surface negativity) is plotted upwards. Stippled lines indicate the 95% confidence interval (± 2 s.e.m.). *d*, Peri-stimulus time histogram of MUA in area 7. Spikes were aligned on the onset of the visual stimulus in the first panel, on the time that the response key was pressed in the second panel, on the rotation of the grating in the third panel, and on the delivery of the food reward in the fourth panel. Trigger spikes used for computation of the field-potential averages in (*a–c*) came from windows that are indicated below the respective panels.

seems likely that these sampling differences account for the observed difference in the strength of correlations revealed by the two measures.

Received 26 June; accepted 19 November 1996.

1. Abeles, M. *Corticonics* (Cambridge Univ. Press, 1991).
2. Singer, W. & Gray, C. M. *Annu. Rev. Neurosci.* **18**, 555–586 (1995).
3. Engel, A. K., König, P., Kreiter, A. K., Schillen, T. B. & Singer, W. *Trends Neurosci.* **15**, 218–226 (1992).
4. Gray, C. M., König, P., Engel, A. K. & Singer, W. *Nature* **338**, 334–337 (1989).
5. Eckhorn, R. et al. *Biol. Cybern.* **60**, 121–130 (1988).
6. Bouyer, J. J., Montaron, M. F. & Rougeul, A. *Electroencephalogr. Clin. Neurophysiol.* **51**, 244–252 (1981).
7. Murthy, V. N. & Fetz, E. E. *Proc. Natl Acad. Sci. USA* **89**, 5670–5674 (1992).
8. Sanes, J. N. & Donoghue, J. P. *Proc. Natl Acad. Sci. USA* **90**, 4470–4474 (1993).
9. Vaadia, E. et al. *Nature* **373**, 515–518 (1995).
10. Bressler, S. L., Coppola, R. & Nakamura, R. *Nature* **366**, 153–156 (1993).
11. Scannell, J. W., Blakemore, C. & Young, M. P. J. *Neurosci.* **15**, 1463–1483 (1993).
12. Felleman, D. J. & Van Essen, D. C. *Cereb. Cortex* **1**, 1–47 (1991).
13. Maunsell, J. H. R. & Newsome, W. T. *Annu. Rev. Neurosci.* **10**, 363–401 (1987).
14. Goodale, M. A. & Milner, A. D. *Trends Neurosci.* **15**, 20–25 (1992).
15. Kalaska, J. F. & Crammond, D. J. *Science* **255**, 1517–1523 (1992).
16. Jeannerod, M., Arbib, M. A., Rizzolatti, G. & Sakata, H. *Trends Neurosci.* **18**, 314–320 (1995).
17. Mitzdorf, U. *Physiol. Rev.* **65**, 37–100 (1985).
18. Engel, A. K., Kreiter, A. K., König, P. & Singer, W. *Proc. Natl Acad. Sci. USA* **88**, 6048–6052 (1991).
19. Fries, A., Eckhorn, R., Bauer, R., Woelbern, T. & Kehr, H. *Neuroreport* **5**, 2273–2277 (1994).
20. Nelson, J. I., Salin, P. A., Munk, M. H. J., Arzi, M. & Bullier, J. *Vis. Neurosci.* **9**, 21–37 (1992).
21. Roelfsema, P. R., Engel, A. K., König, P. & Singer, W. *J. Cogn. Neurosci.* **8**, 603–625 (1996).
22. Engel, A. K., König, P., Kreiter, A. K. & Singer, W. *Science* **252**, 1177–1179 (1991).
23. Munk, M. H. J., Nowak, L. G., Nelson, J. I. & Bullier, J. *J. Neurophysiol.* **74**, 2401–2414 (1995).
24. Buchwald, N. A., Horvath, F. E., Wyers, E. J. & Wakefield, C. *Nature* **201**, 830–831 (1964).
25. Steriade, M., McCormick, D. A. & Sejnowski, T. J. *Science* **262**, 679–685 (1993).
26. Munk, M. H. J., Roelfsema, P. R., König, P., Engel, A. K. & Singer, W. *Science* **272**, 271–274 (1996).
27. Steriade, M., Amzica, F. & Contreras, D. *J. Neurosci.* **16**, 392–417 (1996).
28. Nieoullon, A. & Rispal-Padel, L. *Brain Res.* **105**, 405–422 (1976).
29. Tusa, R. J., Palmer, L. A. & Rosenquist, A. C. *J. Comp. Neurol.* **177**, 213–236 (1978).

ACKNOWLEDGEMENTS. We thank C. Stenner, N. Hensel and S. Herzog for technical assistance.

CORRESPONDENCE and requests for materials should be addressed to P.R.R. (e-mail: roelfsema@mpin-frankfurt.mpg.de).

In vivo dendritic calcium dynamics in neocortical pyramidal neurons

Karel Svoboda, Winfried Denk, David Kleinfeld* & David W. Tank

Biological Computation Research Department, Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA

THE dendrites of mammalian pyramidal neurons contain a rich collection of active conductances that can support Na^+ and Ca^{2+} action potentials (for a review see ref. 1). The presence, site of initiation, and direction of propagation of Na^+ and Ca^{2+} action potentials are, however, controversial², and seem to be sensitive to resting membrane potential, ionic composition, and degree of channel inactivation, and depend on the intensity and pattern of synaptic stimulation. This makes it difficult to extrapolate from *in vitro* experiments to the situation in the intact brain. Here we show that two-photon excitation laser scanning microscopy³ can penetrate the highly scattering tissue of the intact brain. We used this property to measure sensory stimulus-induced dendritic $[\text{Ca}^{2+}]$ dynamics of layer 2/3 pyramidal neurons of the rat primary vibrissa (Sm1) cortex *in vivo*. Simultaneous recordings of intracellular voltage and dendritic $[\text{Ca}^{2+}]$ dynamics during whisker stimulation or current injection showed increases in $[\text{Ca}^{2+}]$ only in coincidence with Na^+ action potentials. The amplitude of these $[\text{Ca}^{2+}]$ transients at a given location was approximately proportional to the number of Na^+ action potentials in a short burst. The amplitude for a given number of action potentials was greatest in the proximal apical dendrite and declined steeply with increasing distance from the soma, with little Ca^{2+} accumulation in the most distal branches, in layer 1.

This suggests that widespread Ca^{2+} action potentials were not generated, and any significant $[\text{Ca}^{2+}]$ increase depends on somatically triggered Na^+ action potentials.

Although *in vivo* experiments provided some of the first evidence for active currents in dendrites⁴, most of our understanding of the presence and role of these dendritic currents comes from *in vitro* brain-slice experiments using electrophysiological recording and optical imaging of $[\text{Ca}^{2+}]$. Such experiments have provided increasing evidence that Na^+ action potentials can, under appropriate conditions, propagate centrifugally into the dendritic tree of cortical pyramidal neurons⁵. Because these action potentials open voltage-dependent Ca^{2+} channels^{6–8} and modulate voltage-dependent synaptic currents⁹, they could be important in the control of long-term synaptic plasticity. Brain-slice experiments have also demonstrated Ca^{2+} action potentials in pyramidal cell dendrites^{10,11} associated with very large dendritic Ca^{2+} accumulations^{6,12}. However, dendritic excitability is strongly affected by a variety of biophysical parameters, such as ionic composition¹³. Furthermore, the site of dendritic action-potential initiation depends on the strength of synaptic excitation^{8,14}, and the activation of voltage-dependent conductances can be modulated by inhibitory synaptic inputs^{15–17}. This implies that it is difficult to extrapolate from *in vitro* experiments to the intact nervous system. Thus understanding the presence and biological function of active dendritic currents requires measurements of membrane voltages¹⁸ and dendritic ion dynamics in the intact nervous system in response to behaviourally relevant stimuli.

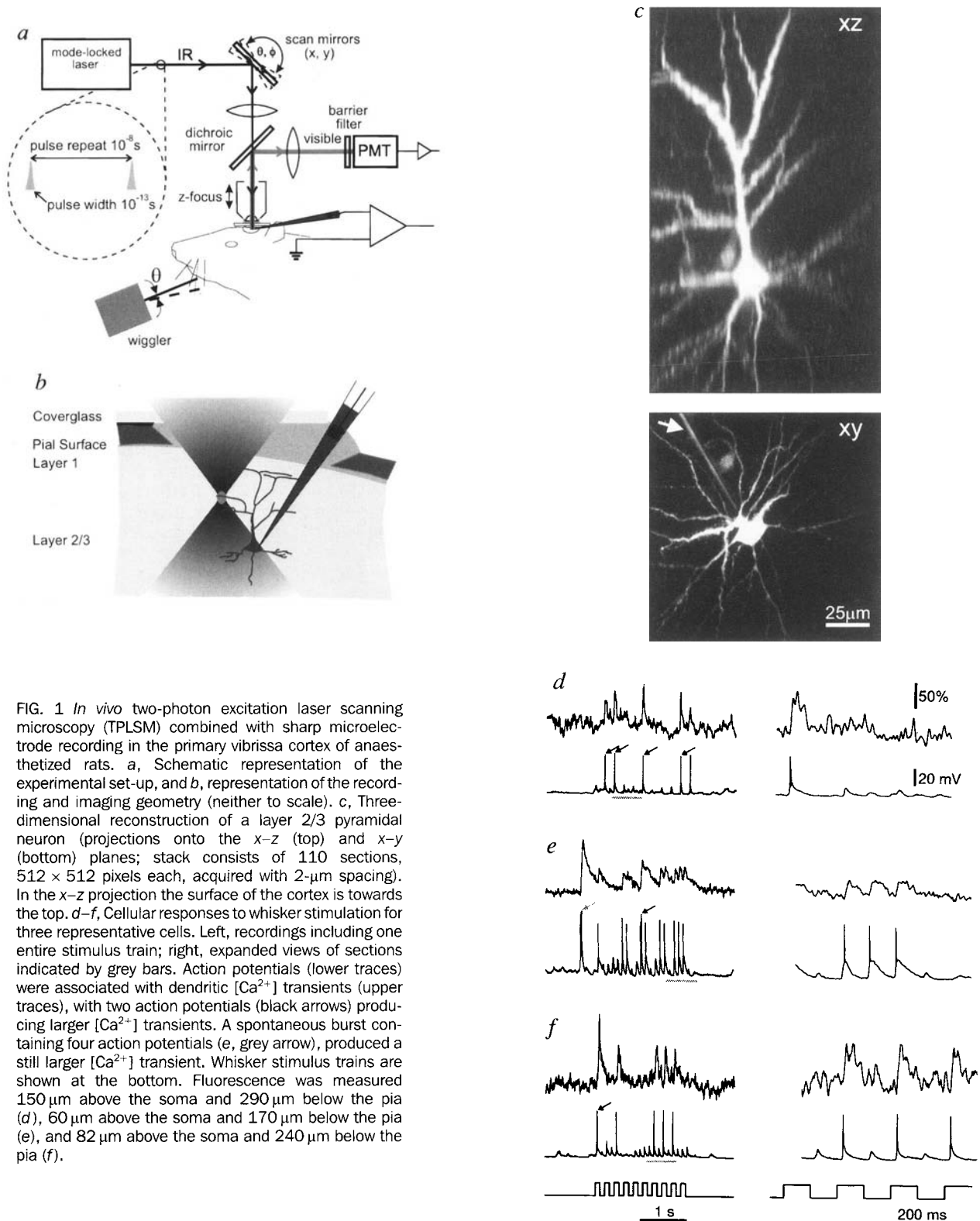
We measured $[\text{Ca}^{2+}]$ dynamics in neocortical neurons of anaesthetized rats by using a two-photon laser scanning microscope (TPLSM)³ that allows imaging while the membrane potential is recorded using a microelectrode (Fig. 1a). In our TPLSM, two infrared photons produced by a mode-locked pulsed laser are simultaneously absorbed to excite fluorophores that normally absorb at visible wavelengths, with an absorption rate proportional to the square of the incident light intensity. Two-photon excitation was important for this experiment for several reasons¹⁹. First, infrared excitation provides greater depth penetration into the cortex than does visible light. Second, excitation is localized to the focal region (excitation volume $\sim 1 \mu\text{m}^3$), providing three-dimensional optical sectioning and resolution equivalent to that of a confocal microscope without any loss of fluorescence light due to a detector pinhole, resulting in much reduced phototoxicity and photobleaching. Regular spiking²⁰ pyramidal cells in layer 2/3 of the primary vibrissa cortex (Sm1) were penetrated with sharp microelectrodes, and were iontophoretically filled with calcium green-1 (Fig. 1b). Dendrites were clearly resolved down to 500 μm below the pial surface, and three-dimensional cell morphology and location of electrode penetration could be reconstructed from stacks of images acquired at different depths (Fig. 1c). For $[\text{Ca}^{2+}]$ measurements at selected locations with high temporal resolution, line-scan mode was used⁹.

Trains of whisker deflections (5 Hz, 2 s) produced a sequence of electrical events consisting of slow depolarizations, during which one, two or no Na^+ action potentials occurred (Fig. 1d–f). All slow depolarizations with Na^+ action potentials were quite large with low amplitude variability (Fig. 1e, f), whereas most that did not contain Na^+ action potentials were smaller and more variable in amplitude. The largest slow depolarizations ranged from 10 mV when recorded in the soma, to 40 mV when recorded in dendrites. Although they resembled large excitatory postsynaptic potentials²¹, the uniformity of slow depolarizations associated with Na^+ action potentials also suggests that regenerative voltage-dependent conductances are involved. Simultaneous microelectrode measurements of membrane potential and fluorescence measurements of calcium indicator-filled dendrites ($n = 49$) revealed $[\text{Ca}^{2+}]$ transients that occurred in one-to-one correspondence with Na^+ action-potential bursts, with more action potentials leading to larger $[\text{Ca}^{2+}]$ transients (Fig. 1d–f, arrows). The transients had fast rise times (~ 2 ms) and short decay-time constants (< 100 ms).

* Present address: Department of Physics 0319, University of California, La Jolla, California 92093, USA.

To examine the possibility that Ca^{2+} current-mediated regenerative events are responsible for the uniformity of slow depolarizations, we examined more closely the $[\text{Ca}^{2+}]$ transient timecourse and the dependence of the $[\text{Ca}^{2+}]$ transient size on the number of action potentials contained in bursts produced by whisker stimulation and depolarizing current pulses. In the proximal dendrites, most of the $[\text{Ca}^{2+}]$ rise occurred within about 2 ms (the instrumental time resolution), and was coincident with the time of the Na^+ action potential, instead of rising gradually throughout the course of the underlying slow event (Fig. 2a–f),

as would be expected for a Ca^{2+} action potential^{6,12}. The size of the $[\text{Ca}^{2+}]$ transient closely followed the number of spikes in a burst (Fig. 2g–i), which is similar to observations in CA1 cells *in vitro*²². No significant influx occurred during large-amplitude slow depolarizations without Na^+ action potentials (Fig. 2a) or subthreshold depolarizing current pulses (Fig. 2d). Injection of hyperpolarizing current during whisker stimulation ($n = 5$) both increased the size of the slow depolarizations and also the fraction without Na^+ action potentials. Slow depolarizations without Na^+ action potentials, were still not accompanied by measurable $[\text{Ca}^{2+}]$ transients



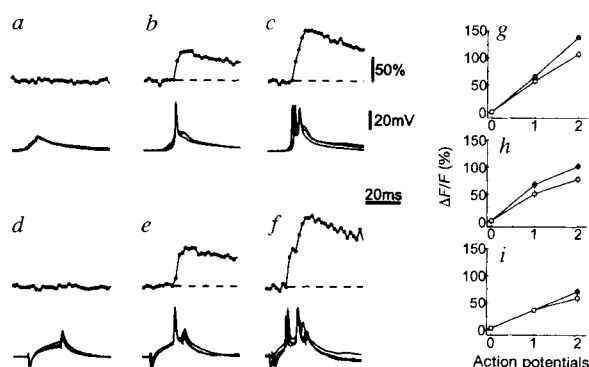


FIG. 2 Representative membrane-potential recordings (lower traces) and spike-triggered averages of $[Ca^{2+}]$ transient amplitude (upper traces; symbol spacing, 2 ms) *in vivo*. $[Ca^{2+}]$ transient and corresponding membrane potential in response to whisker stimulation (a–c) and current injection (d–f) evoking zero (a, $n = 49$; d, $n = 22$), one (b, $n = 76$; e, $n = 27$), or two (c, $n = 57$; f, $n = 28$) action potentials. In c and f averaging was triggered on the second action potential of the pair. g–i, Plots of peak $[Ca^{2+}]$ transient as a function of action-potential number (open circles, whisker stimulation; filled circles, current injection stimulation). g, Apical dendrite (82 μm above the soma, 240 μm below the pia). h, Apical dendrite (150 μm above the soma, 290 μm below the pia). i, Basal dendrite (20 μm below the soma, 180 μm below the pia).

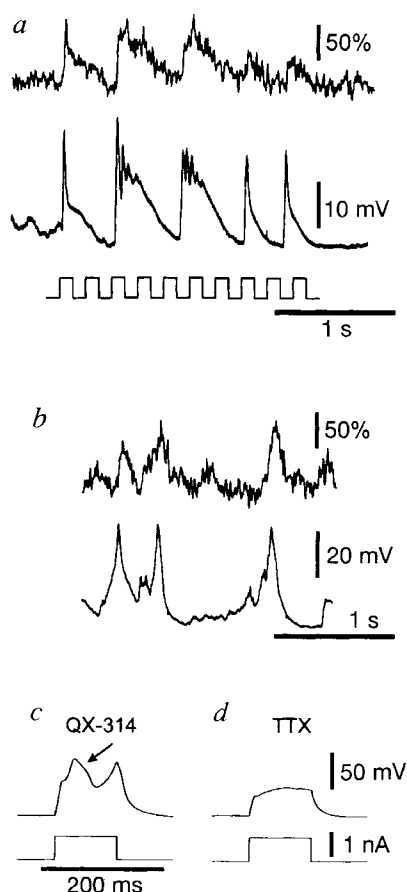


FIG. 3 Effects of QX-314 on electrogenesis *in vivo* (a, b) and *in vitro* (c, d). a, Dendritic $[Ca^{2+}]$ transient (top) and electrical response (middle) to whisker stimulation (bottom); fluorescence measured 50 μm above the soma and 210 μm below the pial surface. b, Spontaneous electrogenesis and associated $[Ca^{2+}]$ transient in a superficial neuron; fluorescence measured 50 μm above the soma and 100 μm below the pial surface. c, Membrane potential (upper trace) in response to a current pulse (lower trace) in a layer 2/3 pyramidal neuron measured *in vitro*, with intracellular QX-314. d, Response in another cell with tetrodotoxin (TTX) in the bath.

(data not shown). We conclude that uniformity of slow depolarizations is not produced by a widespread Ca^{2+} action potential, and that the $[Ca^{2+}]$ transients we observe are produced by the opening of voltage-dependent Ca^{2+} channels during Na^{+} action potentials.

To further test whether Ca^{2+} action potentials were triggered by whisker deflection, we recorded from neurons in which Na^{+} action potentials were abolished by QX-314 ($n = 6$). Dendritic $[Ca^{2+}]$ transients during whisker stimulation now tracked all of the slow depolarizing waveforms, some of which contained oscillations suggesting calcium electrogenesis (Fig. 3a). We also observed large spontaneous membrane-potential excursions (10–30 mV) accompanied by Ca^{2+} influx (Fig. 3b). Further, in neocortical brain slices, intracellular application of QX-314 to layer 2/3 pyramidal neurons produced Ca^{2+} action potentials in response to somatic current injections (Fig. 3c, arrow; $n = 4$), but extracellular application of tetrodotoxin, a highly specific Na^{+} -channel blocker, did not have this effect (Fig. 3d; $n = 4$). Together these results demonstrate that QX-314 assists the generation of Ca^{2+} action potentials *in vivo* (Fig. 3a) and *in vitro* (Fig. 3c), possibly by blocking K^{+} channels^{6,12,23,24}. More importantly, these results show that our imaging technique is sensitive enough to detect Ca^{2+} action potentials if they occur (Fig. 3a, b).

Although our data demonstrate that the uniformity of slow depolarizations is not produced by a widespread Ca^{2+} action potential, we have not ruled out the possibility that it is due to localized Ca^{2+} action potentials in dendritic regions, such as the numerous basal dendrites, that were not sampled adequately in our experiments. However, in slice recordings, dendritic Ca^{2+} action potentials generate the largest and most distributed $[Ca^{2+}]$ transients⁶, so we do not consider localized dendritic Ca^{2+} action potentials to be a likely explanation for our data. An alternative explanation for the uniformity of slow depolarizations is that they are produced by other regenerative currents, which might not cause Ca^{2+} influx, such as the low-voltage-gated Na^{+} current responsible for membrane oscillations in cortical pyramidal cells²⁵. An altogether different explanation may be the generation of uniform synaptic input as a result of some collective network excitation.

The $[Ca^{2+}]$ transient associated with a Na^{+} action potential has been reported to be a uniform global signal in the dendrite⁸. It has also been reported to show a strong spatial dependence²⁶. We examined this issue *in vivo* by mapping the spatial pattern of $[Ca^{2+}]$ transient amplitude in the apical dendrite (Fig. 4a) in response to whisker stimulation (Fig. 4b; $n = 3$), current injection (Fig. 4c; $n = 2$) and extracellular stimulation (Fig. 4d; $n = 1$). For all three stimuli, the $[Ca^{2+}]$ transient amplitude was largest in the proximal apical dendrite, peaking within $\sim 100 \mu\text{m}$ of the soma. Beyond 100 μm the amplitude decreased with distance from the soma, with no detectable $[Ca^{2+}]$ rise in the tufted branches of layer 1 (Fig. 4a–d). The absence of $[Ca^{2+}]$ transients in layer 1 was confirmed in other cells ($n = 7$), without constructing a complete map. Under all stimulating conditions, somatic $[Ca^{2+}]$ transients were clearly present but were small, possibly owing to the unfavourable surface-to-volume ratio, in agreement with *in vitro* measurements^{5,6}. The amplitudes of $[Ca^{2+}]$ transients in basal dendrites in the vicinity of the soma ($< 50 \mu\text{m}$ away) were similar to the largest values found for apical dendrites (Figs 2i and 4b). However, because basal dendrites are thin and deeply embedded in the tissue they were often not bright enough for us to map the $[Ca^{2+}]$ transient amplitude as a function of distance from the soma.

We have demonstrated that TPLSM can be used to measure $[Ca^{2+}]$ dynamics in the dendrites of neurons in the intact mammalian brain. Our measurements render it unlikely that the large-amplitude, low-variability slow depolarizations produced by sensory stimulation in the anaesthetized rat are Ca^{2+} action potentials. The amplitude and spatial pattern of $[Ca^{2+}]$ transients produced by sensory stimulation is dominated by the effects of Na^{+} action-potential depolarization. The fall-off of the $[Ca^{2+}]$

transients with distance, which could be caused by Na^+ action-potential propagation failure or by a non-homogenous distribution of voltage-dependent Ca^{2+} channels^{5,26}, is even steeper than reported in CA1 hippocampal pyramidal cells *in vitro*²⁶, and distinctly different from the unattenuated global $[\text{Ca}^{2+}]$ signal that has also been reported in CA1 cells⁸. The spatial pattern we have observed might imply spatially varying mechanisms or conditions for the induction of Ca^{2+} -dependent long-term synaptic plasticity²⁷ in the highly stratified afferent inputs in neocortex. However, the largest $[\text{Ca}^{2+}]$ transients on the apical dendrite of

layer 2/3 neurons occur in a region devoid of synaptic spines and also, presumably, of excitatory (glutamergic) input. Given the sensitivity of dendritic excitability on K^+ channel modulation^{6,12,24}, it will be interesting to examine whether the pattern we measured is altered by the neuromodulatory influence of cortical activating systems. Our failure to observe Ca^{2+} action potentials in regular spiking layer 2/3 cells does not rule out the presence of regenerative Ca^{2+} events under different stimulus conditions or during different behavioural states. Further experiments using TPLSM might help to address these issues. □

Methods

In vivo imaging and electrophysiology. Rats (4–6 weeks, $n = 50$) were anaesthetized with urethane (2 mg per g body weight, administered intraperitoneally), a small craniotomy (3×3 mm) was cut above primary vibrissa cortex, and the dura was removed. The exposed area was covered with 2% agarose (Type III-A, Sigma, in artificial cerebrospinal fluid) and a coverslide mounted on a stainless-steel frame²⁸ that was secured to the skull with dental cement and fixed to the microscope stage. Sharp electrodes were filled with 400 mM K-acetate and 3 mM calcium green-1 (Molecular Probes) and bevelled (resistance ~ 50 – 100 M Ω). In some cases electrodes also contained 30 mM Lidocaine *N*-ethyl bromide (QX-314, Research Biochemicals International). Membrane-potential recordings were made with an intracellular amplifier (Neurodata IR283) and stored digitally. Neurons with resting potentials more negative than -55 mV were iontophoretically filled with indicator dye by injecting them with hyperpolarizing current (-0.2 to -0.5 nA, 5 to 15 min). During filling, a thin pencil lead attached to a galvanometer scanner was used to deflect one to three whiskers by 5–20 deg in a sequence of ten square deflections at 5 Hz, similar to whisker movement frequencies and amplitudes during tactile discrimination²⁹. Simultaneously, stimulator position was adjusted until a cellular response was detected. No attempt was made to identify the principal whisker. After filling, the electrode was either withdrawn ($n = 18$) or the hyperpolarizing current was turned off ($n = 49$). Cell input resistances ranged from 20 to 120 M Ω , and action-potential amplitudes from 40 to 85 mV. Intracellular recordings lasted between 20 min and 2 h. Electrical recordings with and without ($n = 4$) Ca^{2+} indicator in the electrode were indistinguishable. Imaging was performed with a custom-made laser scanning microscope. The light source for two-photon excitation was a pulsed Ti:sapphire laser (Tsunami; Spectra Physics) operating at 80 MHz repeat frequency, 100 fs pulse width, and wavelength of between 800 and 850 nm. Excitation light was focused by a $\times 40$ water-immersion objective lens (0.75 NA; Carl Zeiss). The average power delivered to the brain was <200 mW. Scanning was performed using two galvanometer mirrors (Cambridge Technology, model 6800). Scanning and image acquisition were controlled with custom software. Emitted light was collected in epifluorescence mode and detected with a photomultiplier tube (Hamamatsu R9638). Acquisition of images and electrophysiology was synchronized to 125 μ s. Fast fluorescence measurements were achieved by repeatedly scanning a single line across the dendrite at 2-ms time intervals. The resulting line-scan image was used for quantitative analysis as follows: the fluorescence signal, F , was computed as an average along the scan line over the width of the dendrite or soma; the baseline signal, F_{base} , was computed as an average over the same width for several scan lines before an event; the normalized change in fluorescence was then $\Delta F/F = (F - F_{\text{base}})/(F_{\text{base}} - B)$, where B is the background signal determined from averages on areas adjacent to the dendrite. Apart from dye saturation, $\Delta F/F$ is proportional to $\Delta[\text{Ca}^{2+}]$. Action potential-evoked $[\text{Ca}^{2+}]$ transients had short decay time constants (<100 ms) (ref. 22). For current injection experiments, current pulses (20 ms) were injected into neurons at 2 Hz. The current amplitude was adjusted to stochastically generate zero, one or two Na^+ action potentials. For both whisker stimulation and current injection stimulation, events with zero, one or two action potentials were grouped, respectively, and spike triggered averages of $\Delta F/F$ were computed. To quantify the size of the $[\text{Ca}^{2+}]$ transients, the peak response was computed as an average over three scan lines around the peak of the response; because of motion-related fluctuations of the fluorescence signal at heartbeat (~ 7 Hz) and breathing frequencies (~ 2 Hz), this measure was found to be more reliable than the integral of the response. The dependence of $\Delta[\text{Ca}^{2+}]$ on action-potential number occasionally contained a slight compressive nonlinearity (see Fig. 2h). Considering that the dendritic $[\text{Ca}^{2+}]$ rise in response to a single Na^+ action potential can approach the K_d of the indicator²², dye saturation may be responsible³⁰, but Na^+ spike waveform changes and Ca^{2+} -channel inactivation may also contribute. Considering that synaptic activation produces localized Ca^{2+} influx *in vitro*⁹, our failure to detect calcium transients in response to subthreshold slow depolarization (Fig. 2a) may seem surprising. But a slow depolarization might be the result of activation of tens of synapses, less than 1% of the total, and our technique samples on the order of 1 μ m of dendrite at any time. It would therefore be very difficult to find Ca^{2+} transients associated with synaptic activation at an unknown location in the dendritic tree⁹. In some mapping experiments during which the electrode was not withdrawn (Fig. 4c), indicator gradients might have reduced $\Delta F/F$ close to the site of the electrode penetration, where the dye concentration is highest. Our

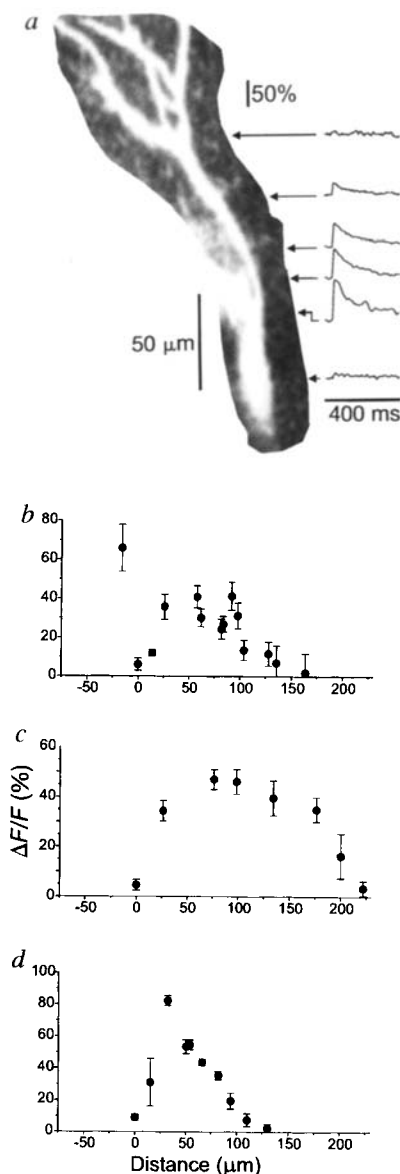


FIG. 4 Spatial maps of average $[\text{Ca}^{2+}]$ transient amplitude in apical dendrites of layer 2/3 neurons *in vivo* using different stimuli. *a*, Layer 2/3 neuron *in vivo* (projections onto the x-z plane; stack consists of 115 sections, 256 $\times$ 256 pixels each, 2 μm spacing; surface of the cortex is towards the top. Right, selected $[\text{Ca}^{2+}]$ transients (average of 5–10 responses). *b*, Whisker deflections (soma 260 μm below the pia). The averaging of $[\text{Ca}^{2+}]$ transients was triggered on extracellularly recorded spikes after electrode withdrawal. *c*, Current injection evoking two Na^+ action potentials (soma was 415 μm below the pia). The averaging of the fluorescence signal was triggered on the second Na^+ action potential. *d*, Extracellular stimulation (1 M Ω monopolar glass electrode filled with 1 M NaCl; 100 μs pulses) on the border of layers 1 and 2 (soma 280 μm below the pia). The averaging of the $[\text{Ca}^{2+}]$ transient was triggered on the stimulus artefact (same cell as in *a*).

basic conclusion, that $[Ca^{2+}]$ transients peak close to the soma with little influx in the distal branches of layer 1, would thus not be affected. In some cases (Fig. 4b, d), to dissipate dye gradients, Ca^{2+} transients were recorded more than 30 min after electrode withdrawal. All data are presented as mean $\pm$ s.e.m.

Brain-slice preparation, imaging and electrophysiology. Rats (4–6 weeks old) were deeply anaesthetized (50 mg ml⁻¹ Nembutal) and decapitated. The brain was removed and coronal slices (300 μ m thick) from somatosensory cortex were cut with a vibratome. Slices were kept at 33 °C for 30 min and subsequently stored at room temperature (22–24 °C). After 1–5 h slices were transferred to a submerged recording chamber (22–24 °C). The artificial cerebrospinal fluid contained (in mM): 124 NaCl, 26 NaHCO₃, 3 KCl, 1.25 NaH₂PO₄, 2 MgSO₄, 10 dextrose, 2 CaCl₂, saturated with 95% O₂/5% CO₂ (pH 7.3; osmolarity, 290–300 mOsm). In some cases 1 μ M of tetrodotoxin was added. Whole-cell recordings were established under visual control, and recordings were made with a patch-clamp amplifier (EPC-7; List Electronics) operating in current-clamp mode. Resistance compensation was performed off-line. Electrodes had resistances of 7–10 M Ω when filled with internal solution containing (in mM): 135 K-methanesulphonate (Fluka), 10 K-HEPES, 2 MgCl₂, 3 Na₂ATP, 0.2 EGTA and 0.1 calcium green-1 (Molecular Probes) (pH, 7.2–7.3; osmolarity, 280 mOsm). In some experiments 6 mM QX-314 was added. Access resistances after break-in were 15–30 M Ω . Measurements were started 20 min after break-in.

Received 5 August; accepted 14 November 1996.

- Yuste, R. & Tank, D. W. *Neuron* **16**, 701–716 (1996).
- Regehr, W. G. & Armstrong, C. M. *Curr. Biol.* **4**, 436–439 (1994).
- Denk, W., Strickler, J. H. & Webb, W. W. *Science* **248**, 73–76 (1990).
- Stuart, G. J. & Sakmann, B. *Nature* **367**, 69–72 (1994).
- Regehr, W. G., Connor, J. A. & Tank, D. W. *Nature* **341**, 533–536 (1989).
- Jaffe, D. B. et al. *Nature* **357**, 244–246 (1992).
- Magee, J. C. & Johnston, D. *Science* **268**, 301–304 (1995).
- Spruston, N., Schiller, Y., Stuart, G. & Sakmann, B. *Science* **268**, 297–300 (1995).
- Yuste, R. & Denk, W. *Nature* **375**, 682–684 (1995).
- Wong, R. K. S., Prince, D. A. & Basbaum, A. I. *Proc. Natl Acad. Sci. USA* **76**, 986–990 (1979).
- Kim, H. G. & Connors, B. W. *J. Neurosci.* **13**, 5301–5311 (1993).
- Yuste, R., Gutnick, M. J., Saar, D., Delaney, K. R. & Tank, D. W. *Neuron* **13**, 23–43 (1994).
- Andreasen, M. & Lambert, J. D. C. *J. Physiol. (Lond.)* **483**, 421–441 (1995).
- Turner, R. W., Meyers, E. R., Richardson, D. L. & Barker, J. L. J. *Neurosci.* **11**, 2270–2280 (1991).
- Kim, H. G., Beierlein, M. & Connors, B. W. *J. Neurophysiol.* **74**, 1810–1814 (1995).
- Buzsaki, G., Penttonen, M., Nadasdy, Z. & Bragin, A. *Proc. Natl Acad. Sci. USA* **93**, 9921–9925 (1996).
- Tsubokawa, H. & Ross, W. N. *J. Neurophysiol.* **76**, 2896–2906 (1996).
- Hirsch, J. A., Alonso, J.-M. & Reid, R. C. *Nature* **378**, 612–616 (1995).
- Denk, W. et al. *J. Neurosci. Methods* **54**, 151–162 (1994).
- Connors, B. W. & Gutnick, M. J. *Trends Neurosci.* **13**, 99–104 (1990).
- Carvell, G. E. & Simons, D. J. *Brain Res.* **448**, 186–191 (1988).
- Helmchen, F., Imoto, K. & Sakmann, B. *Biophys. J.* **70**, 1069–1081 (1996).
- Andreasen, M. & Häblitz, J. J. *J. Neurophysiol.* **69**, 1966–1975 (1993).
- Reuveni, I., Friedman, A., Amitai, Y. & Gutnick, M. J. *J. Neurosci.* **13**, 4609–4621 (1993).
- Gutfreund, Y., Yarom, Y. & Segev, I. *J. Physiol. (Lond.)* **483**, 621–640 (1995).
- Callaway, J. C. & Ross, W. N. *J. Neurophysiol.* **74**, 1395–1403 (1995).
- Bear, M. F. & Malenka, R. C. *Curr. Opin. Neurobiol.* **4**, 389–399 (1994).
- Kleinfeld, D. & Delaney, K. R. *J. Comp. Neurol.* **375**, 89–108 (1996).
- Carvell, G. E. & Simons, D. J. *J. Neurosci.* **10**, 2638–2648 (1990).
- Feller, J. B., Delaney, K. R. & Tank, D. W. *J. Neurophysiol.* **76**, 381–401 (1996).

ACKNOWLEDGEMENTS. We thank R. Stepnoski for writing the software controlling the microscope, and G. Buzsaki, M. Fee and Z. Mainen for discussion.

CORRESPONDENCE and requests for materials should be addressed to W.D. (e-mail: denk@bell-labs.com) or D.W.T. (e-mail: dwt@physics.lucnet.com).

on eumelanin pigment synthesis. The *agouti* peptide is also an antagonist of the hypothalamic melanocortin-4 receptor (MC4-R)^{7–9}. To test the hypothesis that *agouti* causes obesity by antagonism of hypothalamic melanocortin receptors⁷, we identified cyclic melanocortin analogues¹⁰ that are potent agonists or antagonists of the neural MC3 (refs 11, 12) and MC4 receptors. Intracerebroventricular administration of the agonist, MTII, inhibited feeding in four models of hyperphagia: fasted C57BL/6J, *ob/ob*, and *A^y* mice, and mice injected with neuropeptide Y. Co-administration of the specific melanocortin antagonist and *agouti*-mimetic SHU9119 completely blocked this inhibition. Furthermore, administration of SHU9119 significantly enhanced nocturnal feeding, or feeding stimulated by a prior fast. Our data show that melanocortinergic neurons exert a tonic inhibition of feeding behaviour. Chronic disruption of this inhibitory signal is a likely explanation of the *agouti* obesity syndrome.

The *agouti* peptide acts as a paracrine factor in the regulation of eumelanin (brown–black pigment) synthesis, and in the induction of obesity¹³. The non-endocrine nature of the peptide and the lack of homogeneous preparations have complicated analysis of the activities of the peptide *in vivo*. Screening analogues of the cyclic lactam melanocortin agonist MTII (Ac-Nle⁴-c[Asp⁵, D-Phe⁷, Lys¹⁰]- α -MSH-(4-10)-NH₂) led to the identification of the *agouti*-mimetic SHU9119 (Ac-Nle⁴-c[Asp⁵, D-2-Nal⁷, Lys¹⁰]- α -MSH-(4-10)-NH₂)¹⁰ (Fig. 1a). This compound shares pharmacological properties with *agouti* peptide in that it is a potent antagonist of MC4-R and a less potent antagonist of the MC3-R⁷ (Fig. 1b). The closely related analogue MTII (ref. 14) is identical to SHU9119 with the exception of a D-phenylalanine in place of the D-2-naphthylalanine, and is a full agonist of the MC3-R and MC4-R (Fig. 1b).

Mice were induced to feed by food deprivation for 16 h before intracerebroventricular (ICV) administration of the nonspecific melanocortin agonist MTII. In comparison to vehicle-injected animals, MTII was found to produce a potent inhibition of feeding within one hour of administration (Fig. 2a). At the highest dose (3 nmol), food intake was significantly inhibited for up to 4 h after administration ($P < 0.001$), and decreased food intake continued for the next 4 h, with normal rates resuming about 8 h after treatment. There was a dose–response relationship between MTII dose and feeding inhibition, with an IC₅₀ at the 2-h time point of 0.6 nmol. Inhibition of feeding with 3 nmol MTII was blocked by co-administration of 6 nmol SHU9119 (Fig. 2b; $P < 0.001$), demonstrating that the effect results specifically from agonist binding to MC4-R and/or MC3-R.

MTII-treated mice were alert and exhibited no unusual behaviour relative to controls. The effect of MTII on locomotor activity was tested by using sound- and light-proof cages containing multiple light-beam detectors. The movements of MTII- and vehicle-treated mice were measured during two 90-min time periods (0.5–2 h, Fig. 2c; 2.5–4 h, data not shown) after ICV administration. The higher initial activity, indicative of exploratory behaviour, and continued locomotion were indistinguishable between the two groups, indicating that the inhibition of feeding was not due to decreased arousal or locomotion. In contrast to its long-term inhibition of feeding (4–8 h), MTII only inhibited drinking in water-deprived animals during a brief time period (<1 h) after administration (Fig. 2d).

The administration of MTII also inhibited food intake in three other models of hyperphagia, involving C57BL/6J-*Lep^{ob}*, C57BL/6J-*A^y* and neuropeptide Y (NPY)-injected C57BL/6J mice. The hyperphagia in these models can be seen clearly by comparing the 12 h food intake after a fast in vehicle-injected C57BL/6J (2.4 g; Fig. 2a), C57BL/6J-*A^y* (3.7 g, Fig. 3a) and C57BL/6J-*Lep^{ob}* (3.7 g; Fig. 3c) animals. As expected, MTII treatment inhibited food intake after a 16-h fast in the C57BL/6J-*A^y* mouse (Fig. 3a; $P < 0.05$).

MTII, when co-administered with NPY, significantly inhibited the profound stimulation of feeding normally induced by NPY,

Role of melanocortinergic neurons in feeding and the *agouti* obesity syndrome

Wei Fan*, Bruce A. Boston*†, Robert A. Kesterson*, Victor J. Hruby‡ & Roger D. Cone*

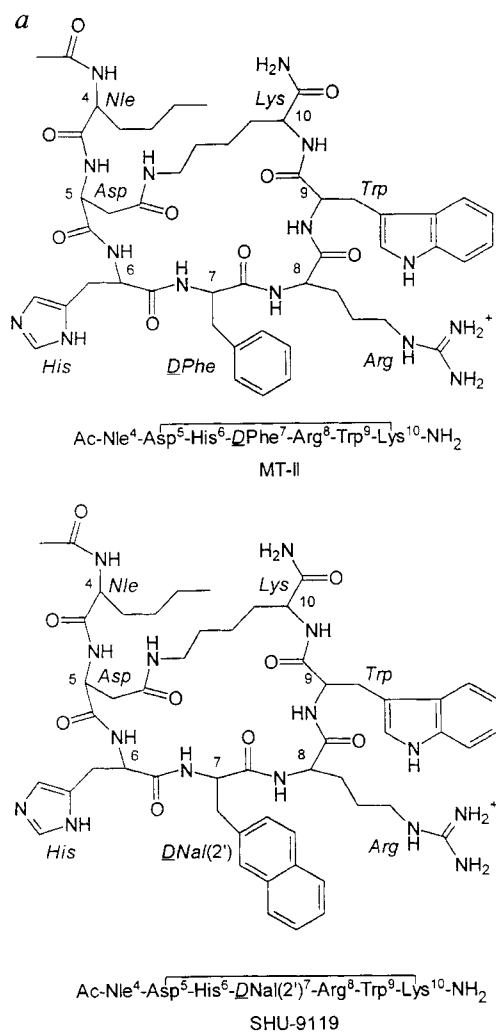
* The Vollum Institute for Advanced Biomedical Research, and

† Department of Pediatrics, Oregon Health Sciences University, 3181 S. W. Sam Jackson Park Road, Portland, Oregon 97201, USA

‡ Department of Chemistry, University of Arizona, Tucson, Arizona 85721, USA

DOMINANT alleles at the *agouti* locus (*A*) cause an obesity syndrome in the mouse, as a consequence of ectopic expression of the *agouti* peptide^{1–6}. This peptide, normally only found in the skin, is a high-affinity antagonist of the melanocyte-stimulating hormone receptor (MC1-R)⁷, thus explaining the inhibitory effect of *agouti*

measured over a 3-h period (Fig. 3b; $P < 0.005$). Co-administration of an approximately twofold molar excess of MTII produced a 74% inhibition of NPY-stimulated food intake at the 3-h time point. Finally, MTII very potently inhibited hyperphagia caused by the absence of leptin in the C57BL/6J-*Lep^{ob}* mouse (Fig. 3c;



b

	rat MC3-R	mouse MC4-R
<i>agouti</i>	antagonist IC ₅₀ >100 nM	antagonist IC ₅₀ =3.9 ± 0.6 nM
SHU9119	antagonist IC ₅₀ =4.5 ± 2.1 nM	antagonist IC ₅₀ =0.36 ± 0.13 nM
MTII	agonist EC ₅₀ =0.27 ± 0.23 nM	agonist EC ₅₀ =0.057 ± 0.024 nM

FIG. 1 Structure and properties of the cyclic lactam melanocortin compounds MTII and SHU9119. **a**, Chemical structures. **b**, Pharmacological properties of MTII and SHU9119 compared with murine *agouti* peptide. IC₅₀ values represent ligand concentration required for half-maximal inhibition of binding of [¹²⁵I]-[Nle⁴, D-Phe⁷]α-MSH tracer. EC₅₀ values represent ligand concentration required for half-maximal activation of a cAMP-responsive β-galactosidase reporter gene. Specific competition of [Nle⁴, D-Phe⁷]α-MSH binding to the rat MC3-R by 100 nM *agouti* peptide was observed, but accurate IC₅₀ values could not be determined in the absence of a peptide preparation of higher concentration and purity.

$P < 0.001$). The C57BL/6J-*Lep^{ob}* mouse was also used to test the ability of MTII to regulate feeding after intraperitoneal administration. Moderate doses (100 nmol) of MTII inhibited feeding in the C57BL/6J-*Lep^{ob}* mouse (Fig. 3d; $P < 0.001$), but low doses (10 nmol) did not (data not shown). The kinetics were similar to those seen with ICV administration, with a potent inhibition of

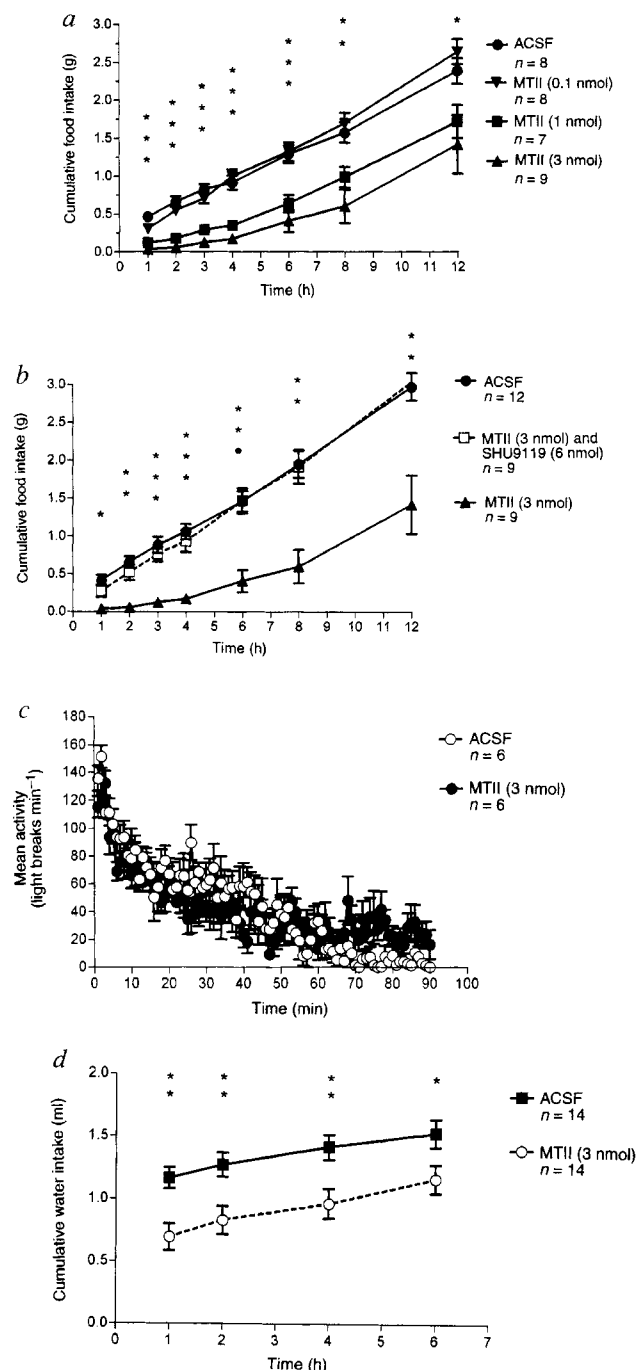


FIG. 2 Inhibition of feeding by ICV administration of the melanocortin agonist MTII. The melanocortin agonist MTII produced a specific and dose-responsive inhibition of feeding in male C57BL/6J mice induced to eat by 16 h food deprivation without altering locomotor activity and with only short-term effects on drinking behaviour. **a**, Inhibition of feeding by MTII is dose responsive. **b**, SHU9119 (6 nmol) specifically blocks MTII inhibition of feeding. **c**, ICV administration of MTII does not alter locomotor activity. **d**, MTII does not produce long-term inhibition of drinking. Significance values indicated for individual time points are: **a**, 3 nmol MTII versus ACSF; **b**, 3 nmol MTII versus 3 nmol MTII + 6 nmol SHU9119. (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$).

feeding for the first 4 h. Melanocortins may also affect metabolism directly because an acute effect of melanocortin agonists on serum insulin could also be seen in the severely hyperinsulinaemic C57BL/6J-*Lep^{ob}* mouse (Fig. 3e), and to a lesser extent in the mildly hyperinsulinaemic C57BL/6J-*A^Y* or normal C57BL/6J mouse (data not shown). Two hours after intraperitoneal administration of [Nle⁴, D-Phe⁷] α -MSH (NDP-MSH), a linear analogue of MTII that has similar pharmacological properties¹⁰, serum insulin levels were 58% lower than in saline-injected controls ($P < 0.001$), and SHU9119 partly blocked the effect ($P < 0.01$); similar but less pronounced effects were seen with MTII.

Daytime food intake in animals fed *ad libitum* was not stimulated by ICV administration of 6 nmol SHU9119 (data not shown). However, SHU9119 induced a significant increase in food intake when administered to animals before lights out (Fig. 4a; $P < 0.02$ at 2 h and 4 h). In this model, 3 nmol SHU9119 led to a mean 29% increase in cumulative food intake at 4 h and a 53% increase in the first 2 h. As anticipated, MTII also significantly inhibited normal nocturnal food intake. Hyperphagia was also induced following SHU9119 treatment of fasted animals (Fig. 4b, $P < 0.001$). Stimulation of feeding was evident at approximately 2 h post-treatment, and continued for 12 h to produce a mean 34% increase in food intake relative to vehicle-injected controls.

Our results demonstrate that a melanocortin agonist can inhibit feeding potentially in four models of hyperphagia, and also inhibit normal nocturnal food intake. Furthermore, an MC3-R/MC4-R antagonist specifically induced hyperphagia, demonstrating that melanocortinergic neurons, probably from the arcuate nucleus^{15,16}, exert a tonic inhibitory effect on feeding behaviour. In a related study, a knockout mouse strain lacking MC4-R has been demonstrated to recapitulate several aspects of the *agouti* obesity syndrome, including hyperphagia, hyperinsulinaemia, late-onset obesity, and increased linear growth¹⁷. Taken together, these results indicate that obesity in the C57BL/6J-*A^Y* mouse results from the cumulative effect of chronic antagonism of MC4-R signalling in the brain by *agouti* peptide. Although it is known that pro-opiomelanocortin mRNA levels in the arcuate nucleus are regulated by metabolic state^{18–20}, the relevant physiological inputs to these neurons, and downstream mechanism by which they regulate feeding and possibly ingestive behaviour in general, remain to be determined.

The elevation of NPY expression in the arcuate nucleus in the C57BL/6J-*Lep^{ob}* mouse²¹ is one downstream factor signalling the induction of hyperphagia in the absence of leptin²²; leptin has been shown to bind to arcuate nucleus neurons and downregulate NPY expression there^{23,24}. No change in arcuate nucleus expres-

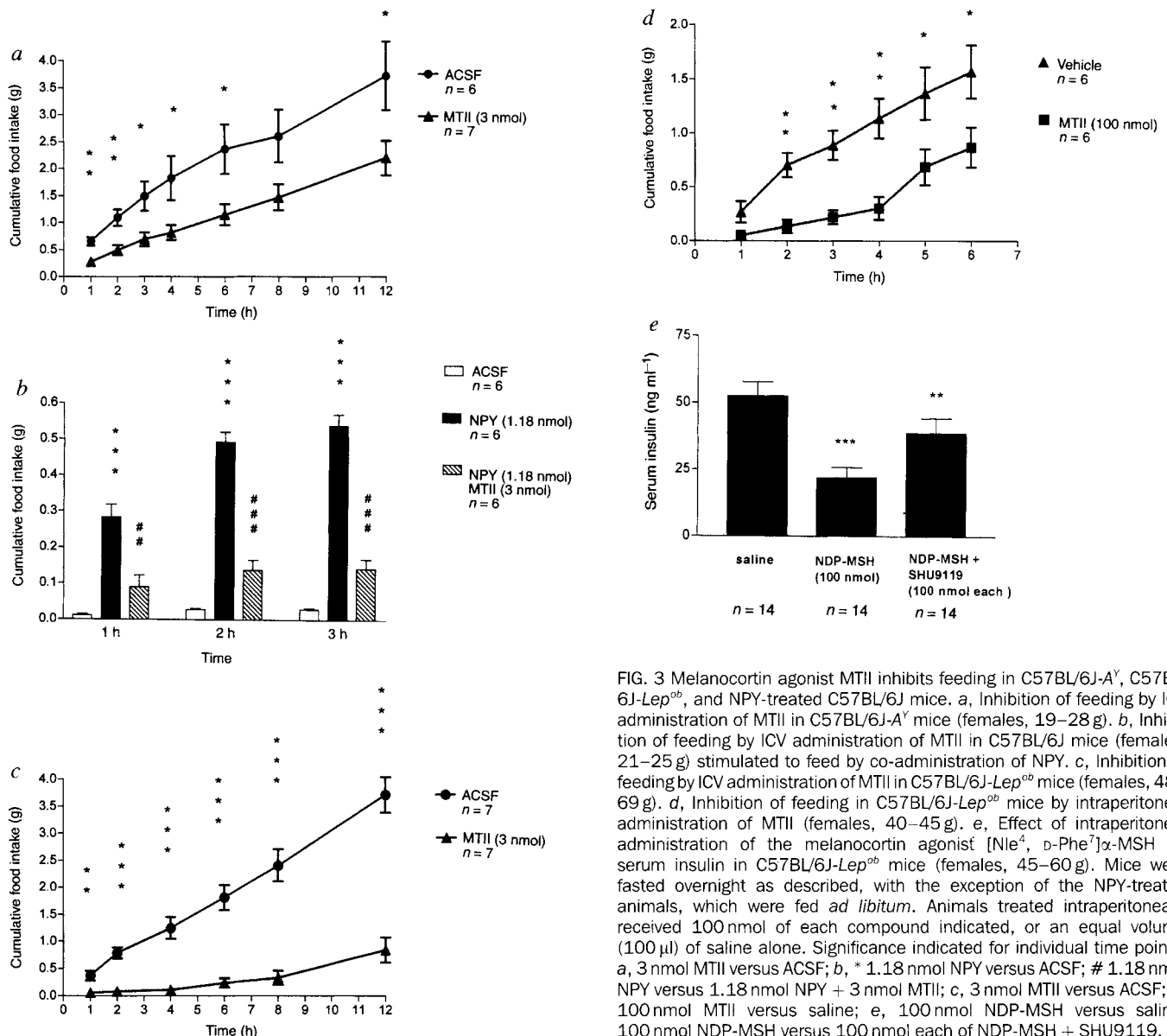


FIG. 3 Melanocortin agonist MTII inhibits feeding in C57BL/6J-*A^Y*, C57BL/6J-*Lep^{ob}*, and NPY-treated C57BL/6J mice. **a**, Inhibition of feeding by ICV administration of MTII in C57BL/6J-*A^Y* mice (females, 19–28 g). **b**, Inhibition of feeding by ICV administration of MTII in C57BL/6J mice (females, 21–25 g) stimulated to feed by co-administration of NPY. **c**, Inhibition of feeding by ICV administration of MTII in C57BL/6J-*Lep^{ob}* mice (females, 48–69 g). **d**, Inhibition of feeding in C57BL/6J-*Lep^{ob}* mice by intraperitoneal administration of MTII (females, 40–45 g). **e**, Effect of intraperitoneal administration of the melanocortin agonist [Nle⁴, D-Phe⁷] α -MSH on serum insulin in C57BL/6J-*Lep^{ob}* mice (females, 45–60 g). Mice were fasted overnight as described, with the exception of the NPY-treated animals, which were fed *ad libitum*. Animals treated intraperitoneally received 100 nmol of each compound indicated, or an equal volume (100 μ l) of saline alone. Significance indicated for individual time points: **a**, 3 nmol MTII versus ACSF; **b**, * 1.18 nmol NPY versus ACSF; # 1.18 nmol NPY versus 1.18 nmol NPY + 3 nmol MTII; **c**, 3 nmol MTII versus ACSF; **d**, 100 nmol MTII versus saline; **e**, 100 nmol NDP-MSH versus saline, 100 nmol NDP-MSH versus 100 nmol each of NDP-MSH + SHU9119.

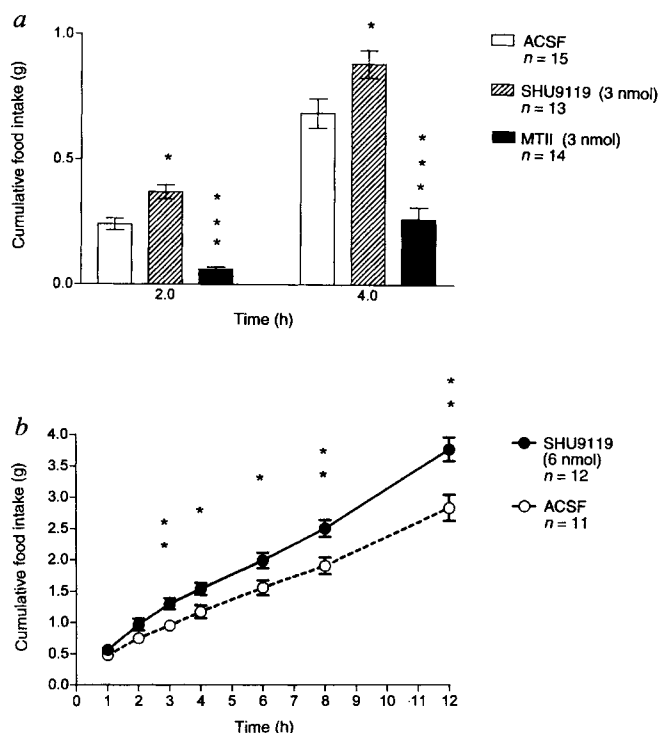


FIG. 4 Stimulation of feeding by ICV administration of the melanocortin antagonist SHU9119. Data show cumulative food intake as a function of time after administration of the compounds shown. *a*, Stimulation of nocturnal feeding with SHU9119 in C57BL/6J mice fed *ad libitum*. *b*, Stimulation of daytime feeding by SHU9119 administration in fasted C57BL/6J mice. Male C57BL/6J mice were maintained on a normal 12 h/12 h light/dark cycle with food (Purina mouse chow) and water *ad libitum*, and treated just before lights off (*a*). In *b*, mice were fasted for 16 h before treatment. Asterisks indicate significance of drug versus control (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$).

sion of NPY mRNA was seen in C57BL/6J- A^Y mice. However, NPY is induced in the dorsomedial hypothalamic nucleus in obese C57BL/6J- A^Y and MC4-R knockout mice, which also identifies NPY as a potential downstream effector in these two melanocortinergic obesity models²⁵. □

Methods

Peptide preparation and analysis. SHU9119 and MTII were synthesized and purified to homogeneity^{10,14}. Murine *agouti* peptide was produced in the baculovirus system, as ref. 7 but with one modification: *agouti* was purified from baculovirus supernatants by 0.6-M NaCl step elution from an EconoS cation exchange column (BioRad); peptide was approximately 60% pure. Agonist and antagonist activities of MTII, SHU9119 and *agouti* were first ascertained by examining their ability to stimulate the adenylyl cyclase signalling pathway in HEK293 cell lines expressing the indicated receptor (rat MC3-R and mouse MC4-R), or to inhibit stimulation by α -MSH. A direct adenylyl cyclase assay²⁶ was used to analyse *agouti*, but MTII and SHU9119 were analysed using a cAMP-responsive β -galactosidase reporter construct²⁷ to detect intracellular cAMP levels. Competition binding experiments were performed¹⁰, with nonspecific binding determined as the amount of radioactivity bound in the presence of 5×10^{-6} M cold [Nle⁴, D-Phe⁷] α -MSH, and was typically 3–5% of total counts bound. Values shown are the mean and standard deviations from 2–3 determinations.

Injections and feeding assays. Mice (20–30 g male C57BL/6J, unless indicated otherwise) were maintained on a normal 12 h/12 h light/dark cycle with food (Purina mouse chow) and water *ad libitum*. Animals were housed individually for 24 h, distributed into weight-matched experimental and control groups, and injected with vehicle or vehicle plus drug as indicated. Fasting consisted of food deprivation from 18:00 to 10:30 to stimulate feeding during the daytime experimental period. Animals were lightly anaesthetized with halothane, and administered buffered artificial cerebrospinal fluid (ACSF, in

mM) (137.9 NaCl, 3.37 KCl, 1.5 CaCl₂, 1 MgCl₂, 1.45 NaH₂PO₄ · H₂O, 4.85 Na₂HPO₄ · 7H₂O, pH 7.4) as vehicle, or compounds indicated in a volume of 2 μ l ACSF into one lateral ventricle. Freehand ICV injections were performed essentially as described²⁸. A high efficiency of injection into the lateral ventricle was achieved, as monitored using methylene blue, in training sessions before the experiments were started. The compounds indicated, in a 2 μ l volume of ACSF, were administered slowly over approximately 15 s, and the needle removed after 35 s. Mice were allowed to recover from anaesthesia and placed into a cage containing a premeasured quantity of food pellets in a spill-free cup. Only one moribund animal and no fatalities resulted from the 206 animals receiving ICV injections in this study; this animal and its matched pair were excluded. Food remaining was briefly removed and weighed at the time intervals indicated. Data points indicate the mean, bars indicate standard error. Significance of effects over time was determined by analysis of variance (ANOVA) with repeated measures. Significance of drug effects at individual time points was determined by two-way ANOVA, and is indicated in each figure (* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$).

Locomotor assay. Locomotor assays were performed by first injecting 3 nmol MTII or ACSF into mice, 3 h after lights on, with mice on food and water *ad libitum*. At 0.5 h after the injection, 12 mice were placed without prior habituation into separate boxes containing multiple infrared light sources and photodetectors. Locomotor activity was then monitored for two 90-min periods, 0.5–2 and 2.5–4 h after injection. The boxes were contained within separate ventilated light- and sound-attenuating chambers (Coulbourn model E10-20). Disruption of the infrared beams, with 10 ms resolution, was tallied independently for each 1-min time period in each cage. Data points indicate the mean total activity (number of light breaks) for 6 animals in each experimental group, and bars indicate standard error. Four-way ANOVA was used to analyse the data, and indicated that there was no significant effect of drug on locomotor activity.

Drinking assay. Mice on food *ad libitum* were deprived of water for 24 h before the experiment. They were then injected ICV with ACSF or MTII (3 nmol) and housed individually in cages containing volumetric water bottles but no food. Water consumption was measured hourly.

Insulin assay. C57BL/6J-*Lep^{ob}* mice (female, 45–60 g) were injected intraperitoneally 1 h after lights on with 100 μ l saline, 100 nmol NDP-MSH, or 100 nmol of each NDP-MSH + SHU9119. After 2.5 h, 50 μ l blood was withdrawn by tail bleed, and the serum separated and assayed (10 μ l) for insulin, in duplicate, by radioimmunoassay (Linco).

Received 2 August; accepted 21 November 1996.

- Klebig, M. L., Wilkinson, J. E., Geisler, J. G. & Woychik, R. P. *Proc. Natl Acad. Sci. USA* **92**, 4728–4732 (1995).
- Bultman, S. J., Michaud, E. J. & Woychik, R. P. *Cell* **71**, 1195–1204 (1992).
- Kucera, G. T., Bortner, D. M. & Rosenberg, M. P. *Dev. Biol.* **173**, 162–173 (1996).
- Perry, W. L., Hustad, C. M., Swing, D. A., Jenkins, N. A. & Copeland, N. G. *Genetics* **140**, 267–274 (1995).
- Wilson, B. D. et al. *Hum. Mol. Genet.* **4**, 223–230 (1995).
- Miller, M. W. et al. *Genes Dev.* **7**, 454–467 (1993).
- Lu, D. et al. *Nature* **371**, 799–802 (1994).
- Gantz, I. et al. *J. Biol. Chem.* **268**, 15174–15179 (1993).
- Mountjoy, K. G., Mortrud, M. T., Low, M. J., Simerly, R. B. & Cone, R. D. *Mol. Endocrinol.* **8**, 1298–1308 (1994).
- Hruby, V. J. et al. *J. Med. Chem.* **38**, 3454–3461 (1995).
- Gantz, I. et al. *J. Biol. Chem.* **268**, 8246–8250 (1993).
- Roselli-Rehffuss, L. et al. *Proc. Natl Acad. Sci. USA* **90**, 8856–8860 (1993).
- Wolff, G. L. *Genetics* **48**, 1041–1058 (1963).
- Al-Obeidi, F., Castrucci, A. M., Hadley, M. E. & Hruby, V. J. *J. Med. Chem.* **32**, 2555–2561 (1989).
- Watson, S. J., Akil, H., Richard, C. W. & Barchas, J. D. *Nature* **275**, 226–228 (1978).
- Jacobowitz, D. M. & O'Donohue, T. L. *Proc. Natl Acad. Sci. USA* **75**, 6300–6304 (1978).
- Huszar, D. et al. *Cell* (in the press).
- Bergendahl, M., Wiemann, J. N., Clifton, D. K., Huhtaniemi, I. & Steiner, R. A. *Neuroendocrinology* **56**, 913–920 (1992).
- Brady, L. S., Smith, M. A., Gold, P. W. & Herkenham, M. *Neuroendocrinology* **52**, 441–447 (1990).
- Steiner, R. A., Kabigting, E., Lent, K. & Clifton, D. K. *J. Neuroendocrinol.* **6**, 603–608 (1994).
- Wilding, J. P. H. et al. *Endocrinology* **132**, 1939–1944 (1993).
- Erickson, J. C., Holloper, G. & Palmiter, R. D. *Science* **274**, 1704–1707 (1996).
- Stephens, T. W. et al. *Nature* **377**, 530–532 (1995).
- Schwartz, M. W. et al. *J. Clin. Invest.* **98**, 1101–1106 (1996).
- Kesterson, R. A., Huszar, D., Lynch, C. A., Simerly, R. S. & Cone, R. D. *Mol. Endocrinol.* (submitted).
- Johnson, R. A. & Salomon, Y. *Methods Enzymol.* **195**, 3–21 (1991).
- Chen, W., Shields, T. S., Stork, P. J. S. & Cone, R. D. *Anal. Biochem.* **226**, 349–354 (1995).
- Laursen, S. E. & Belknap, J. K. *J. Pharmacol. Methods* **16**, 355–357 (1986).

ACKNOWLEDGEMENTS. We thank W. Yuan for synthesis and purification of SHU9119; C. Cunningham for help with locomotor assays; D. Hatten, A. Belknap, D. Huszar, F. Lee and A. Campfield for discussions; and L. Gearhart and D. Horne for help with veterinary care. This work was funded by the National Institutes of Health (NIDDK to R.D.C., V.J.H., B.A.B. and R.A.K., and NICHD to R.D.C.), and Millenium Pharmaceuticals, Inc. and Northwest Neurologic, Inc.

CORRESPONDENCE and requests for materials should be addressed to R.D.C. (e-mail: cone@ohsu.edu).

Phosphotyrosine-dependent activation of Rac-1 GDP/GTP exchange by the vav proto-oncogene product

Piero Crespo^{*†}, Kornel E. Schuebel[‡], Amy A. Ostrom[‡], J. Silvio Gutkind^{*} & Xosé R. Bustelo[‡]

^{*} Molecular Signaling Unit, Laboratory of Cellular Development and Oncology, National Institute of Dental Research, National Institutes of Health, Bethesda, Maryland 20892, USA

[†] Departamento de Biología Molecular, Facultad de Medicina, Universidad de Cantabria, Santander 39011, Cantabria, Spain

[‡] Department of Pathology, School of Medicine and University Hospital, State University of New York at Stony Brook, Stony Brook, New York 11794, USA

THE oncogenic protein Vav^{1,2} harbours a complex array of structural motifs, including leucine-rich, Dbl-homology, pleckstrin-homology, zinc-finger, SH2 and SH3 domains. Upon stimulation by antigens or mitogens, Vav becomes phosphorylated on key tyrosine residues³⁻⁵ and associates with other signalling proteins, including the mitogen receptors^{3,4} Zap-70 (ref. 6), Vap-1 (ref. 5) and SIp-76 (ref. 7). Disruption of the *vav* locus by homologous recombination causes severe defects in signalling by primary antigen receptors, leading to abnormal lymphocyte proliferation and lymphopenia^{8,9}. Despite the importance of Vav cell signalling, the function of this protein remains unknown. Here we show that tyrosine-phosphorylated Vav, but not the non-phosphorylated protein, catalyses GDP/GTP exchange on Rac-1, a protein implicated in cell proliferation and cytoskeletal organization^{10,11}, causing this GTPase to switch from its inactive to its active state. Transfection experiments also show that phosphorylation of Vav on tyrosine residues leads to nucleotide exchange on Rac-1 *in vivo* and stimulates c-Jun kinase, a downstream element in the signalling pathway involving this GTPase. Our results have identified a function for Vav and define a mechanism in which engaged membrane receptors activate its signalling pathway.

We have previously shown that Vav is located upstream of Rac-1 (ref. 12). Because Vav contains a DH (Dbl-homology) domain usually present in guanine-nucleotide-exchange factors (GEFs) that act for GTPases of the Rho/Rac family¹¹, we investigated whether Vav had similar enzyme activity towards Rac-1. Initial efforts using either full-length, oncogenic Vav protein or isolated domains were unsuccessful (data not shown). As Vav becomes extensively phosphorylated on tyrosine residues upon activation of protein tyrosine kinases³⁻⁵, we investigated whether this post-translational modification could influence Vav GDP/GTP-exchange activity. A hexahistidine-tagged Vav protein was incubated with ATP in the presence or absence of a glutathione *S*-transferase (GST) protein containing the protein tyrosine kinase Lck, a member of the Src family¹³; the GDP/GTP exchange activity of Vav was then assessed by following the incorporation of [³⁵S]GTP-γS into GDP-loaded Rac-1 in a filter immobilization assay. Figure 1a shows that preincubation of Vav with GST-Lck fusion protein stimulated the exchange activity of Vav towards Rac-1 protein. Non-phosphorylated Vav caused little GDP/GTP exchange even after 45 min at room temperature, with results being the same as those for Rac-1 and the GST-Lck fusion protein alone (Fig. 1a, left). The kinetics of Rac-1 GDP/GTP exchange induced by phosphorylated Vav were similar to those obtained with Cdc42 and its bonafide GEF, the *dbl* oncogene product¹¹, under the same conditions (Fig. 1a, right). However, unlike Vav, incubation with GST-Lck did not affect the catalytic activity of Dbl (Fig. 1a, right), so upregulation of Vav by this protein tyrosine kinase is not a general feature of all DH family members. We also analysed the effect of Vav on the incorporation

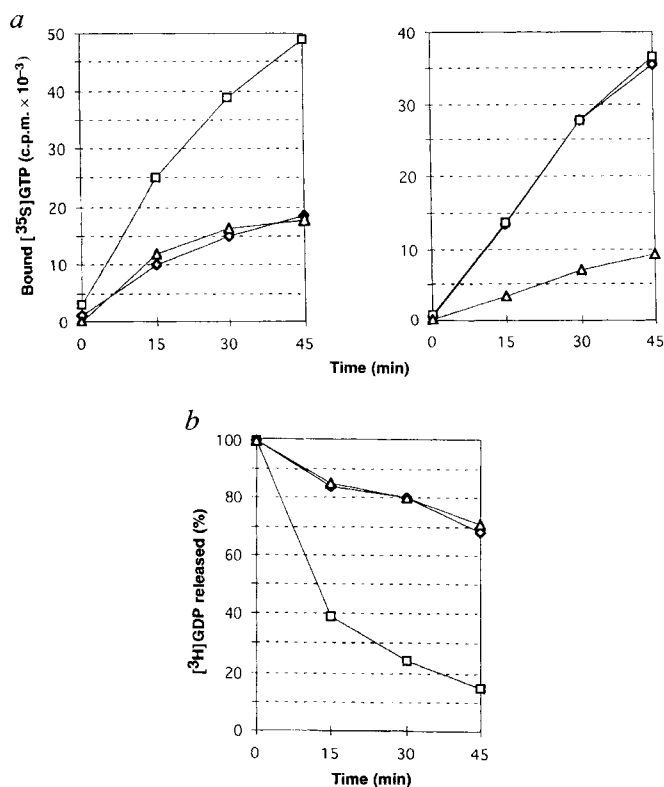


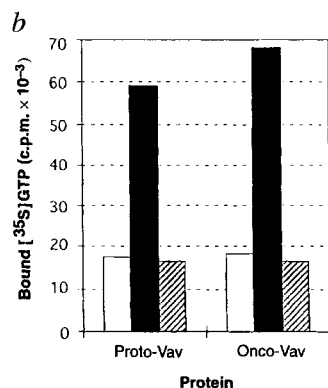
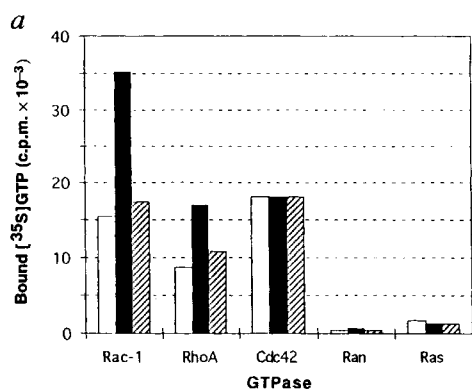
FIG. 1 a, Incorporation of [³⁵S]GTP-γS in Rac-1 mediated by Vav protein. Left, GDP-loaded Rac-1 was incubated with 5 μM [³⁵S]GTP-γS in the presence of phosphorylated Vav (squares), non-phosphorylated Vav (diamonds), or autophosphorylated GST-Lck (triangles). At each time point, aliquots were taken in duplicate and the GTP bound to Rac-1 was evaluated in a filter-immobilization assay. Right, GDP-loaded Cdc42 was subjected to exchange reactions in the presence of either Dbl plus GST-Lck (squares), Dbl alone (diamonds), or autophosphorylated GST-Lck (triangles). b, Kinetics of Rac-1 [³H]GDP release in the presence of 1 mM GTP and the combination of proteins indicated in a (left).

of cold GTP into [³H]GDP-loaded Rac-1 (Fig. 1b) and found that when it is tyrosine-phosphorylated, Vav enhances the dissociation of [³H]GDP from Rac-1 compared with either non-phosphorylated Vav or autophosphorylated GST-Lck protein.

We investigated the specificity of this enzyme by GDP/GTP exchange assay on Ras-family substrates. Phosphorylated Vav was less effective at stimulating nucleotide exchange on RhoA protein than on Rac-1 under the same conditions (Fig. 2a). There was no significant effect when Cdc42, Ran or H-Ras were included in the incubations (Fig. 2a). The lack of Vav exchange activity on H-Ras is in agreement with earlier results indicating that *vav* transformation is independent of Ras function¹⁴. We conclude that Vav is acting *in vitro* as a Rac-1 GEF and that this activity is modulated by tyrosine phosphorylation.

The Vav protein is oncogenically activated as a result of an amino-terminal deletion (residues 1–66). This truncated protein is 20-fold more active than the wild type, as determined by focus formation assay in rodent fibroblasts². To determine whether truncation results in constitutively active Vav, we compared its exchange activity with full-length Vav and found that both proteins were dependent on phosphorylation for their activation (Fig. 2b). This suggests that tyrosine phosphorylation *per se* cannot account for the different activity of the proto- and oncogenic Vav proteins *in vivo*. Activation of Vav *in vivo* might also depend on the elimination of an inhibitory component constitutively associated with its N terminus, the nature of which is under investigation.

To confirm the role of Vav as a Rac-1 GEF, we tested whether expression of the Vav oncoprotein could promote guanosine-



nucleotide exchange on Rac-1 in living cells. Because Rho/Rac proteins have high intrinsic GTPase activity which precludes detection of their GTP-bound forms *in vivo*¹⁵, we used ³²P-labelled GDP bound to these proteins to evaluate their nucleotide-exchange rates¹⁶. AU5-epitope-tagged Rac-1, RhoA, Cdc42 or H-Ras were expressed in COS-7 cells either alone or in combination with oncogenic Vav; cells were transfected, starved, incubated briefly with [³²P]orthophosphate, and nucleotide exchange was determined by counting the GDP in anti-AU5 immunoprecipitates after thin-layer chromatography¹⁶. Co-transfection of the *vav* oncogene with AU5-Rac-1 caused a sevenfold increase in ³²P-labelled GDP bound to Rac-1, but there was no exchange upon co-transfection with AU5-H-Ras or AU5-RhoA; a small exchange was detected for AU5-Cdc42 (Fig. 3a, top). As before¹⁶, incorporation of labelled GDP on these GTPases was low when transfected on their own so they probably undergo minimal nucleotide exchange in quiescent conditions (Fig. 3a). Western blotting confirmed that all proteins were expressed in each transfection (Fig. 3a, bottom panel). As an additional control, nucleotide exchange was routinely activated on RhoA and Cdc42 upon co-transfection with Dbl, and on H-Ras with SOS or Ras GRF/Cdc25^{Mm} (ref. 11) (data not shown), demonstrating that these GTPases are active in our system.

Coexpression of the *vav* oncogene with Rac-1 N17, a dominant-negative mutant that is blocked in the inactive, GDP-loaded form, gave no incorporation of ³²P-labelled nucleotide in Rac-1, con-

FIG. 2 a, Specificity of Vav enzyme activity. Autophosphorylated GST-Lck (hatched boxes), phosphorylated Vav (black boxes), and non-phosphorylated Vav (white boxes) were subjected to exchange reactions in the presence of the indicated GDP-loaded GTPases for 45 min (see Methods). b, Phosphorylation-dependent exchange activity of wild-type and oncogenic Vav proteins. GDP-loaded Rac-1 was incubated with [³⁵S]GTP-γS in the presence of GST-Lck (hatched boxes), phosphorylated (black boxes) or non-phosphorylated (white boxes) Vav proteins. After 45 min, duplicate aliquots were used to determine ³⁵S-labelled GTP incorporated in Rac-1 (see Methods).

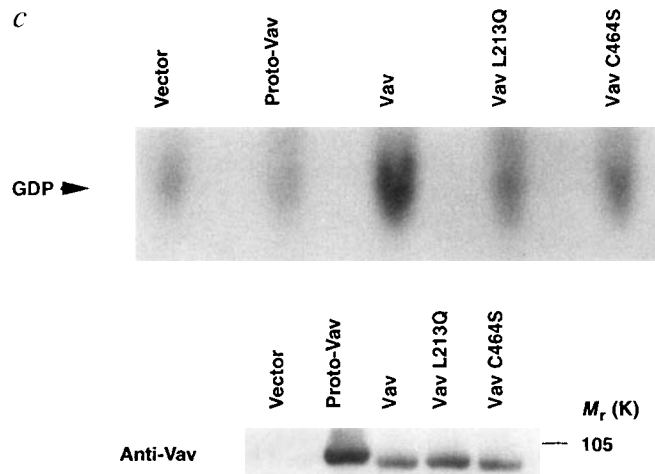
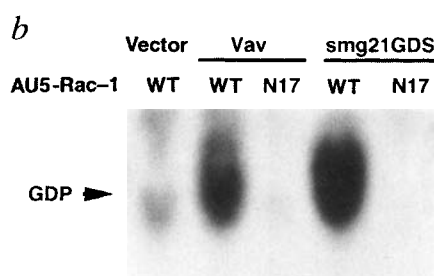
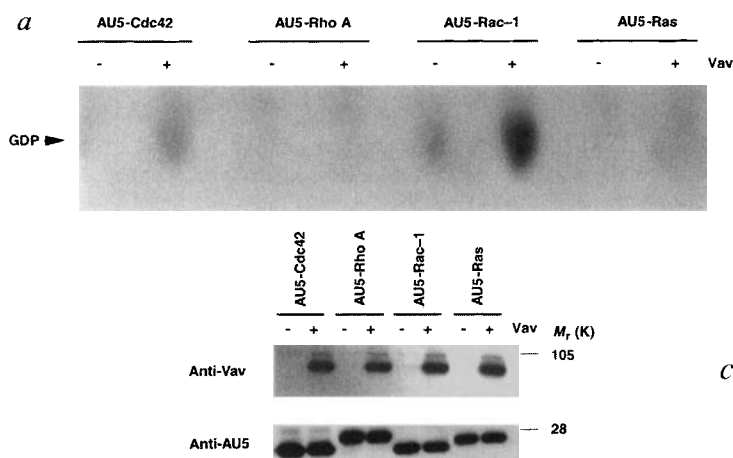


FIG. 3 a, Effect of the *vav* oncogene product on Ras-family nucleotide exchange *in vivo*. Top, nucleotide loading of GTPases in the absence (-) or presence (+) of the *vav* oncogene product. Bottom, expression of Vav and GTPases in COS-7 cells, as determined by anti-Vav⁵ and anti-AU5 (Babco) immunoblot analysis. b, Effect of Vav oncoprotein and smg21GDS on the nucleotide loading of wild type (WT) and a dominant-negative mutant (N17) version of Rac-1. c, Activation of Rac-1 by Vav proteins *in vivo*. Top, nucleotide loading on Rac-1 in the presence of the *vav* proto-oncogene (Proto-Vav), the *vav* oncogene (Vav), and two inactive mutant versions of the *vav* oncogene. Bottom, expression of Vav proteins in COS-7 cells, as determined by anti-Vav immunoblot analysis.

firming that the effect observed in the Vav/Rac-1 co-transfection did reflect an increase in guanine-nucleotide-exchange rate (Fig. 3b). This effect is identical to that of the smg21 guanine-nucleotide-dissociation factor¹¹ (Fig. 3b). Anti-AU5 immunoblot analysis of cell lysates from parallel transfections showed that both forms of Rac-1 were comparably expressed (data not shown). In addition, co-transfection of wild-type Rac-1 with either the *vav* proto-oncogene product or two inactive versions of the Vav oncoprotein with point mutations in the DH motif (L213Q) (ref. 12) or in the cysteine-rich region (C464S) (refs 2, 12) resulted in much slower nucleotide exchange on Rac-1 *in vivo* (Fig. 3c). This agrees with the low activity of these proteins in both focus-formation² and c-Jun kinase (JNK) activation experiments¹² and supports the idea of a linear pathway between Vav, Rac-1 and JNK proteins.

To test whether the activity of the *vav* proto-oncogene product was enhanced by tyrosine phosphorylation, we determined the nucleotide-exchange rate of Rac-1 *in vivo* after co-transfection of the *vav* proto-oncogene with plasmids encoding wild-type or constitutively active mutants of two Src-family members, Lck and Fyn (ref. 13). Although expression of each of these signalling proteins alone did not cause Rac-1 nucleotide exchange, coexpression of the *vav* proto-oncogene with either form of Lck or Fyn synergistically increased ³²P-nucleotide incorporation on Rac-1 (Fig. 4a).

We next examined whether coexpression of the *vav* proto-oncogene product (proto-Vav) with Lck and Fyn allowed stimulation of JNK, a downstream element of the Vav pathway that is activated in a Rac-1-dependent manner¹². A haemagglutinin (HA)-tagged version of JNK was transfected in COS-7 cells either alone or in combination with proto-Vav and wild-type and constitutively active mutants of Lck and Fyn. Expression of Lck and Fyn alone failed to activate JNK, and expression of proto-Vav alone only weakly stimulated JNK (Fig. 4b); in contrast, coexpression of proto-Vav with either Lck or Fyn proteins increased JNK activity (3- to 4-fold), reaching ~60% of levels obtained upon expression of the *vav* oncogene (Fig. 4b). Both nucleotide exchange on Rac-1 and JNK activation upon co-transfection with Lck and Fyn correlated with the induction of tyrosine phosphorylation on proto-Vav under those transfection

conditions, as determined by anti-phosphotyrosine immunoblot analysis of Vav immunocomplexes (Fig. 4c).

Our observations provide evidence for a role for Vav in Rac-1 activation and a mechanisms for stimulating this function during signal transduction by mitogenic and antigenic receptors. According to this model, extracellular stimulation results in phosphorylation of Vav on tyrosine residues and activation of its exchange activity towards Rac-1 and of its downstream elements, including the SEK/JNK serine/threonine cascade. To our knowledge, this is the first example of regulation of a small G protein by direct tyrosine phosphorylation of its GEF. Rac-1 and Vav have been implicated in several overlapping cellular functions, including mitogenesis^{1,2,17,18}, cooperativity with Ras^{14,17,18}, and regulation of NF-AT-dependent responses in T-cell lines^{19,20}. We have shown previously that Rac-1 dominant-negative mutants, but not those of RhoA, Cdc42 or H-Ras, downregulate Vav signalling¹². Considered with our latest results, this indicates that Rac-1 is an important effector molecule in the signalling cascade of the *vav* proto-oncogene product. □

Methods

Protein purification. Wild-type Vav was purified from Sf9 cells by using nickel columns (Qiagen) according to the manufacturer's recommendation. Vav oncoprotein was purified from Sf9 cell lysates using protein-A Sepharose beads coated with a polyclonal anti-Vav antibody raised against residues 576–589 of mouse Vav protein², followed by elution with saturating amounts of immunizing peptide. GST-Lck was purified from Sf9 as described²¹. A polyglutamic-tagged version of the *dbl* oncogene product purified from Sf9 cells was a gift from R. A. Cerione. GTPases were purified as GST fusion proteins according to standard procedures.

Expression vectors. Mammalian expression vectors encoding Vav (pcDNA derivatives) and GTPases (pCEV derivatives) have been described^{12,22}. Constitutively active mutants of Src family proteins, including a truncated version of wild-type Fyn lacking the C-terminal 15 amino acids (FynC), as well as the Lck Y505F mutant (LckC), were expressed using pcDNA-derived vectors (Invitrogen).

Kinase and exchange reactions. Kinase reactions were done for 20–30 min at room temperature in 20 mM Tris-HCl, pH 7.5, 5 mM MgCl₂, 1 mM ATP. [³²S]GTP-γS was incorporated at room temperature²³ using 15 pmol GTPase, 4.5 pmol of either Vav or Dbl, and 3.5 pmol GST-Lck. [³H]GDP release experi-

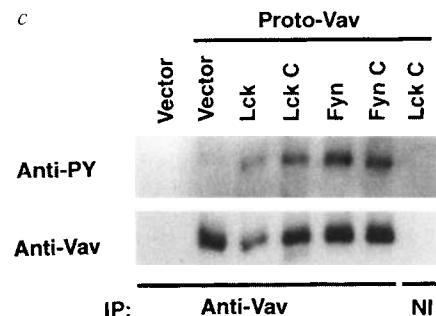
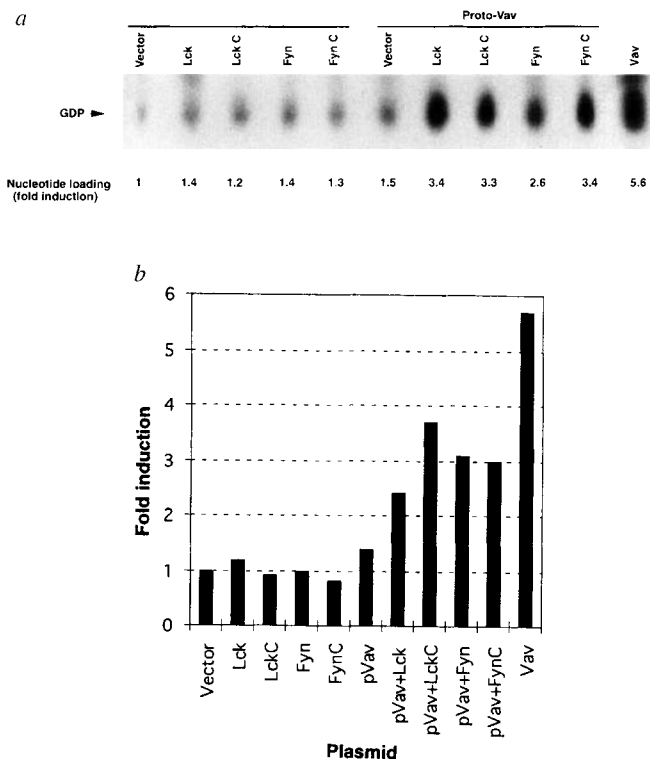


FIG. 4 a, Stimulation of Rac-1 nucleotide exchange rate *in vivo* by tyrosine phosphorylation of the *vav* proto-oncogene product. AU5-tagged Rac-1 was transfected in COS-7 cells either in the absence (–) or presence (Proto-Vav) of the *vav* proto-oncogene and wild-type Lck, Fyn and constitutively active (C suffix) mutants of the indicated Src family members. For comparative purposes, AU5-tagged Rac-1 was coexpressed with Vav. After transfection, relative guanosine-nucleotide-exchange rates on Rac-1 were determined as described in Methods. Values represent the mean of three independent experiments. b, Effect of tyrosine phosphorylation of wild-type Vav (pVav) on JNK activity. Cells were transfected and JNK activity determined using an immunocomplex kinase assay. Values represent the mean of three independent experiments. c, Levels of tyrosine phosphorylation of wild-type Vav in the presence of Src-family kinases. Lysates derived from COS-7 expressing the indicated signalling molecules were immunoprecipitated (IP) with either non-immune serum (NI) or a specific anti-Vav antibody. Final immunocomplexes were analysed on anti-phosphotyrosine (PY) or anti-Vav immunoblot.

ments were done as described²⁴ using 34 pmol GTPase, 3.4 pmol of either Vav or Dbl, and, when appropriate, 3.5 pmol GST-Lck.

JNK assay and *in vivo* nucleotide labelling of GTPases. COS-7 cells were transfected using DEAE-dextran¹². After transfection, cells were cultured for 48 h, serum-starved in phosphate-free DMEM medium for 18 h, labelled with [³²P]orthophosphate (100 µCi ml⁻¹) for 2 h, and disrupted in lysis buffer (50 mM Tris-HCl, pH 7.5, 20 mM MgCl₂, 150 mM NaCl, 0.5% Nonidet-P40, 1 mM sodium orthovanadate, 1 mM PMSF, 25 µg ml⁻¹ leupeptin and 25 µg ml⁻¹ aprotinin). Lysates were immunoprecipitated with an anti-AU5 monoclonal antibody (Babco) for 1 h and immunocomplexes recovered using Gamma-bind G-Sepharose beads (Pharmacia/LKB). Immunoprecipitates were washed three times in lysis buffer, twice in 50 mM Tris-HCl, pH 7.5, 20 mM MgCl₂, 500 mM NaCl, and finally resuspended in 1 M KH₂PO₄, 5 mM EDTA, pH 8.0. Bound nucleotides were released by heating, and fractionated using polyethyleneimine thin-layer chromatography plates¹⁶. JNK activity was determined using an *in vitro* immunocomplex assay^{12,22}.

Received 24 September; accepted 18 November 1996.

1. Katrav, S., Martin-Zanca, D. & Barbacid, M. *EMBO J.* **8**, 2283–2290 (1989).
2. Coppola, J., Bryant, S., Koda, T., Conway, D. & Barbacid, M. *Cell Growth Differ.* **2**, 95–105 (1991).
3. Bustelo, X. R., Ledbetter, J. F. & Barbacid, M. *Nature* **356**, 68–71 (1992).
4. Margolis, B. *et al.* *Nature* **356**, 71–74 (1992).
5. Bustelo, X. R. & Barbacid, M. *Science* **256**, 1196–1199 (1992).

6. Katrav, S., Sutherland, M., Packham, G., Yi, T. & Weiss, A. *J. Biol. Chem.* **269**, 32579–32585 (1994).
7. Wu, J., Motto, D. G., Koretzky, G. A. & Weiss, A. *Immunity* **4**, 593–602 (1996).
8. Zhang, R., Alt, F. W., Davidson, L., Orkin, S. H. & Swat, W. *Nature* **374**, 470–473 (1995).
9. Tarakhovskii, A. *et al.* *Nature* **374**, 467–470 (1995).
10. Ridley, A. J. *Curr. Opin. Genet. Dev.* **5**, 24–30 (1995).
11. Boguski, M. S. & McCormick, F. *Nature* **366**, 643–654 (1993).
12. Crespo, P. *et al.* *Oncogene* **13**, 455–460 (1996).
13. Bolen, J. B. *Cell Growth Differ.* **2**, 409–414 (1991).
14. Bustelo, X. R., Suen, K. L., Leftheris, K., Meyers, C. A. & Barbacid, M. *Oncogene* **9**, 2405–2413 (1994).
15. Zheng, Y., Hart, M. J. & Cerione, R. A. *Methods Enzymol.* **256**, 77–84 (1995).
16. Laudanna, C., Campbell, J. J. & Butcher, E. C. *Science* **271**, 981–983 (1996).
17. Qiu, R. G., Chen, J., Kim, D., McCormick, F. & Symons, M. *Nature* **374**, 457–459 (1995).
18. Qiu, R. G., Chen, J., McCormick, F. & Symons, M. *Proc. Natl Acad. Sci. USA* **92**, 11781–11785 (1995).
19. Wu, J., Katrav, S. & Weiss, A. *Mol. Cell. Biol.* **15**, 4337–4346 (1995).
20. Genot, E., Cleverley, S., Henning, S. & Cantrell, D. *EMBO J.* **15**, 3923–3933 (1996).
21. Spana, C., O'Rourke, E. C., Bolen, J. B. & Fagnoli, J. *Protein Expr. Purif.* **4**, 390–397 (1993).
22. Coso, O. A. *et al.* *Cell* **81**, 1137–1146 (1995).
23. Horii, Y., Beeler, J. F., Sakaguchi, K., Tachibana, M. & Miki, T. *EMBO J.* **13**, 4776–4786 (1994).
24. Shou, C., Farnsworth, C. L., Neel, B. G. & Feig, L. A. *Nature* **358**, 351–354 (1992).

ACKNOWLEDGEMENTS. We thank the Bristol-Myers Squibb Pharmaceutical Research Institute, and M. Barbacid in particular, for providing most of the reagents used in this study. We also thank J. Glaven and R. A. Cerione for purified Dbl, Y. Takai for smg21 GDS cDNA, and N. Reich and M. Dosil for their help and support.

CORRESPONDENCE and requests for materials should be addressed to X.R.B. (e-mail: xbustelo@path.som.sunysb.edu).

Unusual Rel-like architecture in the DNA-binding domain of the transcription factor NFATc

Scot A. Wolfe*, Pei Zhou*, Volker Dötsch†, Lin Chen*, Angie You*, Steffan N. Ho‡, Gerald R. Crabtree‡, Gerhard Wagner† & Gregory L. Verdine*

* Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA

† Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA

‡ Departments of Developmental Biology and Pathology, Howard Hughes Medical Institute, Stanford University Medical School, Stanford, California 94305, USA

TRANSCRIPTION factors of the NFAT family regulate the production of effector proteins that coordinate the immune response¹. The immunosuppressive drugs FK506 and cyclosporin A (CsA) act by blocking a Ca²⁺-mediated signalling pathway leading to NFAT. Although FK506 and CsA have enabled human organs to be transplanted routinely, the toxic side-effects of these drugs limit their usage. This toxicity might be absent in antagonists that target NFAT directly. As a first step in the structure-based search for NFAT antagonists, we now report the identification and solution structure of a 20K domain of NFATc (NFATc-DBD) that is both necessary and sufficient to bind DNA and activate transcription cooperatively. Although the overall fold of the NFATc DNA-binding domain is related to that of NF-κB p50 (refs 2, 3), the two proteins use significantly different strategies for DNA recognition. On the basis of these results, we present a model for the cooperative complex formed between NFAT and the mitogenic transcription factor AP-1 on the interleukin-2 enhancer.

NFAT consists of two clearly separable functional domains. The N-terminal domain of NFATc, residues 1 to 415, controls its subcellular localization in response to Ca²⁺-mobilization (C. R. Beals and G.R.C., manuscript submitted). Residues 416 to 591 of NFATc (NFATc-DBD), are necessary and sufficient for DNA binding and transcriptional activation on the distal antigen-receptor response element of the interleukin (IL)-2 enhancer

(ARRE2)⁴ (S.N.H. and G.R.C., unpublished results). *In vitro*, NFATc-DBD cooperates with AP-1 to form a tight, oriented complex on ARRE2 (ref. 5). Furthermore, NFATc-DBD alone binds specifically to ARRE2, albeit more weakly than full-length NFATc. NFATc-DBD thus represents the core functional domain of human NFATc, similar to that identified for murine NFATp (ref. 6).

The core structural motif of NFATc-DBD consists of a ten-stranded antiparallel β-barrel (Fig. 1a, b) assembled by the packing of three β-sheets. The two primary sheets (βIHCFe and βABG) form the core of the β-barrel, with the third sheet (βDG') capping off one end. Devoid of α-helical structure, the remainder of NFATc-DBD is composed of loops, two of which are especially prominent (βA–βB and βG'–βH loops). As commonly observed for surface projections in NMR structures, the conformations of the two prominent loops in NFATc-DBD vary widely among the family of calculated structures shown in Fig. 1a. The origin of this behaviour, revealed through ¹⁵N-relaxation measurements⁷ (data not shown), was found to differ for the two loops. Whereas the βG'–βH loop has high conformational mobility throughout, the majority of the βA–βB loop is conformationally constrained, except at residues 28–35 near the tip (Fig. 1c), which appear to be only partially ordered.

NFAT family members share extensive sequence homology (~70%) over a ~300-amino-acid region comprising the entire DNA-binding domain, the N-terminal half of which corresponds to NFATc-DBD⁸. A potential sequence relationship between the DNA-binding domain of NFAT and members of the NF-κB/Rel protein family has been noted^{9,10}; however, the extent of sequence similarity (<20%) falls below that considered to be statistically significant. Comparison of the solution structure of NFATc-DBD with the X-ray structure of NF-κB p50 DNA-binding domain (Fig. 2) reveals that the two share an essentially identical overall fold (Fig. 2a, b) and strand topology (Fig. 2c, d). Many of the residues that are invariant between NFAT family members and p50 (Fig. 1c, bold) form the hydrophobic core of the β-barrel; the packing interface between the βAB and βHI strands is nearly identical in the two classes of proteins. The sequence is also highly conserved throughout much of the βG' and βH strands, because both faces of the small βG'H sheet are important: side chains on one face project inwards to establish the protein core, whereas those from the other face (Ile 123, Val 147) make stabilizing contacts to the lower end of the βA–βB loop (Val 38). Several other invariant residues, especially prolines 21, 46, 58, 110 and 156 in NFAT, are

FIG. 1 *a, b*, Solution structure of NFATc-DBD. *a*, Backbone (N, C α , C') overlay of ten calculated structures of NFATc-DBD. The core domain consists of residues 13–172; residues 1–12 and 173–178 are highly disordered. The ten β -strands that comprise the β -barrel, as well as the two prominent loops, are colour-coded. *b*, MOLSCRIPT representation of the NFATc-DBD structure, viewed from the same perspective as in *a* and with the same colour coding. The highly mobile β G'– β H loop is shown in red as a family of C α traces. Constituent strands that make up the three β -sheets of the structure are denoted accordingly in matching colours. Asterisk denotes a five-residue insert (residues 77–81) in the C-terminal end of the p50 DBD, which is without counterpart in the NFATc DBL (asterisk in Fig. 2*b*); this projection contains several basic residues that may contact phosphates outside the canonical DNA recognition site. *c*, Sequence and strand alignment of the four known NFAT isoforms^{8–10,25,26} and human NF- κ B p50 (refs 27, 28). β -strands are denoted above the sequence, colour-coded as in *a* and *b*. The numbering above the sequence corresponds to that of NFATc-DBD, which corresponds to that of human NFATc (ref. 10) minus 413. The italicized *MK* in the sequence of NFATc-DBD denotes amino acids added for purposes of overexpression, which are not present in NFATc itself.

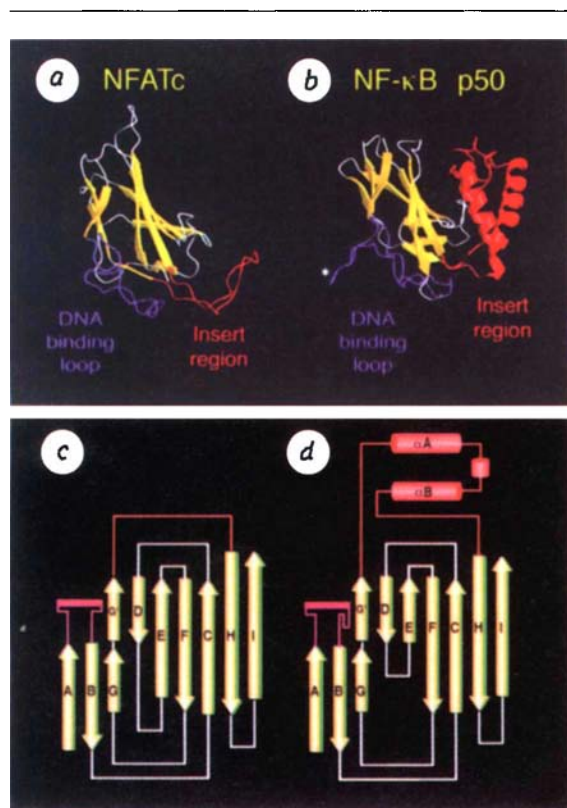
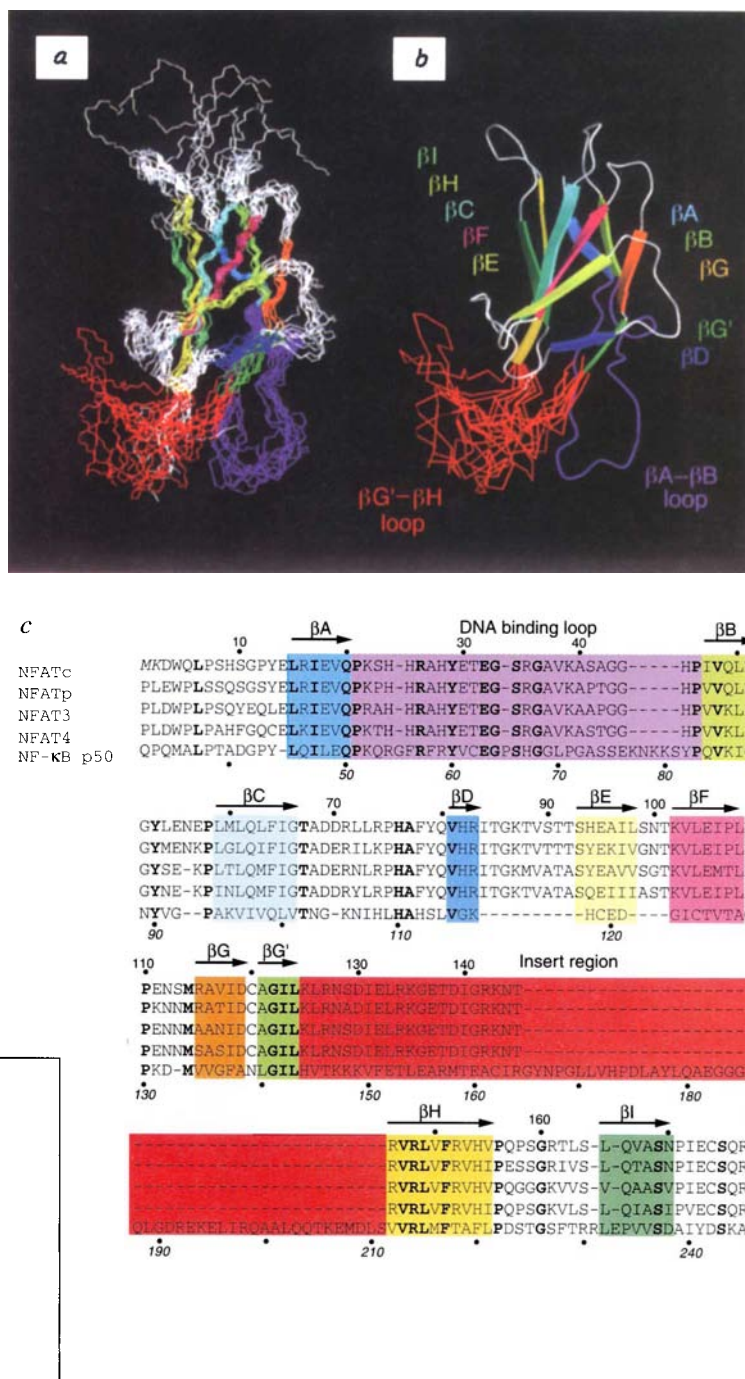


FIG. 2 *a, b*, Comparison of the solution structure (*a*) of NFATc-DBD with the X-ray crystal structure (*b*) of the DNA-binding domain of NF- κ B p50 (ref. 2). Yellow, β -strands; magenta, DNA-binding loop (DBL); red, insert region. Asterisk denotes a basic pentapeptide in the p50 DBL with no counterpart in NFATc-DBD (refer also to Fig. 1*c*); this segment in p50 is believed to make nonspecific contacts with the DNA backbone. *c, d*, Topological maps of the DNA-binding domain of NFATc-DBD (*c*) and the DNA-binding domain of NF- κ B p50 (*d*). β -strands are shown as arrows, α -helices as cylinders. Colour scheme in *c, d* corresponds to that in *a, b*.

important in defining the junction between β -strands and their connecting loops. Many residues of the β A- β B loop are also identical in NFAT and p50 (see below).

The two structures do show significant differences, but these lie not in the strands that make up the β -barrel, but rather in the length and conformation of the segments connecting these strands. Most striking are the differences in the unusually long connector located between strands β G' and β H, which in Rel

proteins is known as the 'insert region'. In NF- κ B p50, the 67 residues of the insert region form a compact helical bundle (Fig. 2*b*, red) which packs tightly against the immunoglobulin-like β -barrel^{2,3}, whereas the much smaller insert region of NFATc-DBD (21 residues) projects out into solution and is unstructured.

The sequence similarity between NFAT and Rel proteins is high throughout the β A- β B loop (residues 20-46 of NFATc-DBD; Fig. 1*c*). All but one of the sequence-specific contacts made

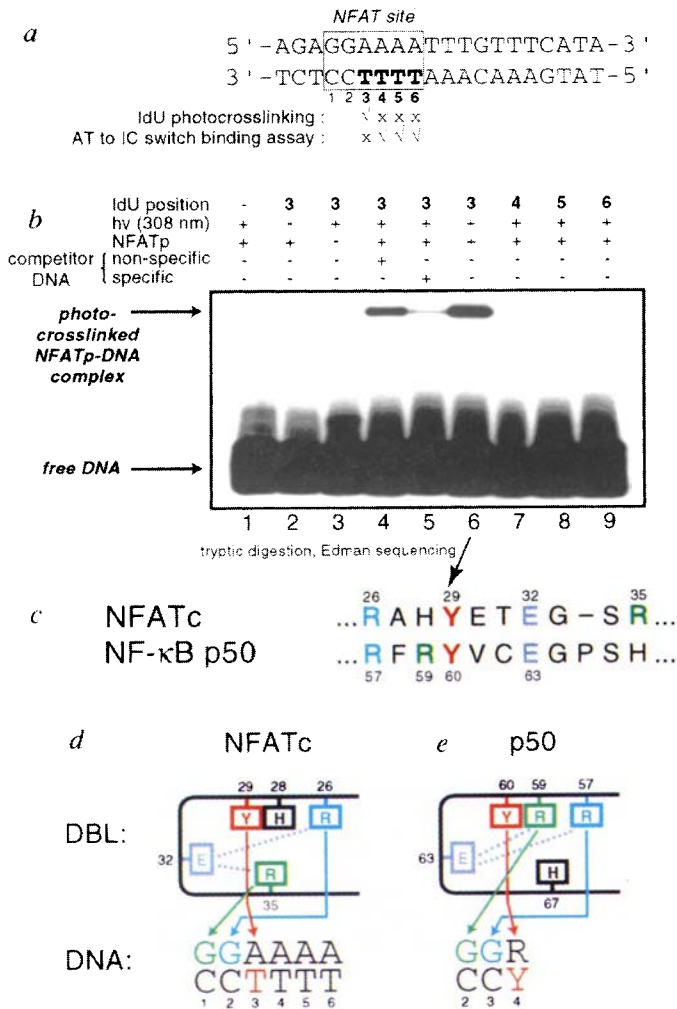


FIG. 3 *a*, Sequences of the ARRE2 oligonucleotides. Variants of the sequence shown containing individual dT residues substituted by 5-iododeoxyuridine (IdU) were used for *b*; variants having A-T base pairs switched to I-C were used in EMSA binding assays as a test for major-groove interactions. A tick indicates that photocrosslinking or binding ($K_d < 20$ nM) was significant; a cross indicates a lack of photocrosslinking or binding ($K_d > 300$ nM). *b*, Photocrosslinking of NFATp to ARRE2. 32 P end-labelled duplex oligonucleotides (top) containing individual IdU residues at the indicated positions within the NFAT site (see *a*) were irradiated with 308-nm UV light under various conditions. The products were separated by denaturing SDS-PAGE (below) and detected by phosphorimaging. A large-scale photocrosslinking reaction (lane 6) led to the identification of a covalent linkage between Tyr29 and position 3, which is ordinarily occupied by a T (data not shown). *c*, Alignment of residues in the p50 DBL responsible for sequence-specific recognition of DNA (57-67) with the corresponding residues in NFATc (26-35). *d*, *e*, Diagram comparing the predicted major-groove contact from the DBL of NFATc-DBD (*d*) with those found in the X-ray structure of p50 (refs 2, 3) (*e*). Relevant portions of the respective binding sites are shown below the DBLs; arrows denote contact pairs. Bridging hydrogen bonds between a key conserved Glu (Glu32 in NFATc-DBD; Glu63 in p50) and Arg residues involved in DNA recognition are indicated by dotted lines. The red contact pair has been determined for both NFATp (*b*, *c*) and p50 (ref. 11) by photocrosslinking, and confirmed in the case of p50 by X-ray crystallography^{2,3}.

TABLE 1 Structural statistics for NFATc-DBD

Number of amino acids assigned (non-proline sequential assignments)	156 of 167
Total number of distance constraints	1,539
Number of effective distance constraints	1,087
Intramolecular	309
Sequential	321
Medium-range	71
Long-range	386
Number of dihedral angle constraints	331
Distance constraint violations >0.2 Å (per structure)	1.57 ± 1.31
Dihedral constraint violations $>2.5^\circ$ (per structure)	1.66 ± 1.21
X-PLOR potential energy (E_{total} , average per structure)	388 ± 15
R.m.s.d. to the mean for backbone heavy atoms, excluding poorly constrained regions (residues 13-26; 36-124; 144-172)	1.17 ± 0.22
R.m.s.d. to the mean for heavy atoms, excluding poorly constrained regions (residues 13-26; 36-124; 144-172)	1.60 ± 0.26
R.m.s.d. to the mean for backbone heavy atoms of all β -strands (residues 15-19; 47-51; 59-66; 81-83; 93-97; 102-108; 115-123; 146-155; 165-170)	0.70 ± 0.07
R.m.s.d. to the mean for heavy atoms of all β -strands	1.26 ± 0.11

All ϕ , ψ angles in the calculated structures lie within allowed regions of the Ramachandran plot.

by p50 involve residues on this DNA-binding loop (DBL)^{2,3}, and mutations within the corresponding segment of NFATp abrogate DNA binding⁶. To identify conclusively the segment of NFATp that contacts DNA in the major groove, we carried out photocrosslinking experiments using oligonucleotides substituted with 5-iododeoxyuridine (IdU) at various single sites within ARRE2 (Fig. 3a). Strong ultraviolet-dependent crosslinking occurred only when IdU was present at position T3 (Fig. 3b). Edman sequencing of the DNA-bound peptide derived from the tryptic digestion revealed that the peptide was crosslinked to DNA through the residue corresponding to Tyr 29 of NFATc-DBD (Fig. 3c). The analogous residue in p50, Tyr 60, undergoes highly efficient BrdU-mediated photocrosslinking to the fourth position of the NF- κ B recognition site (Fig. 3c, e)¹¹. In the crystal structure of p50, the aromatic ring of Tyr 60 is in direct van der Waals contact with the pyrimidine base to which it becomes photocrosslinked^{2,3}. Taken together, these findings indicate that the aromatic ring of Tyr 29 in the NFATc DBL contacts T3 of its DNA-recognition site (Fig. 3d). Knowledge of this conserved contact pair thus permits alignment of the DBLs of NFATc and p50 on their respective DNA sites, and permits prediction of the major-groove DNA contacts used by NFATc (Fig. 3d, e).

Both the NFAT and p50 recognition sites contain a virtually invariant GG dinucleotide immediately adjacent to the cross-linked base pair (Fig. 3d, e). p50 recognizes these G residues through hydrogen-bonding interactions involving Arg 57 and Arg 59 (refs 2, 3). The former Arg is conserved in NFAT proteins (Arg 26 of NFATc-DBD), whereas the latter is replaced by a His (His 28) (Fig. 3c); Arg 26 is essential for DNA binding, whereas His 28 is not⁶. Nonetheless, the overall structure of the DBL in NFATc-DBD is similar to that of p50 (Fig. 2a, b), several key DNA contact residues in p50 are conserved in NFAT (Fig. 3b), and the critical interaction involving Tyr 29/60 is maintained in both. It is therefore likely that the DBLs of p50 and NFAT have similar DNA-bound conformations and make several conserved contacts, in addition to that involving Tyr 29. Thus, the side chain of His 28 in NFATc-DBD would be too short to contact G1, yet interference footprinting indicates that NFAT recognizes this guanine¹². This issue was explained by modelling experiments, which suggested that the side chain of Arg 35 should ideally be positioned to contact G1. We thus propose that the DBLs of NFAT and p50 share a common scaffold for DNA recognition, and present the same side-chain functionality to the 5' GGR sequence, but differ in the disposition of these residues along the scaffold.

Our analysis defines the sequence-specific interactions between NFAT and the 5'-half site of its recognition element (5'-GGAAAA), but leave unanswered the basis for specific recognition of the 3'-half site (5'-GGAAAA). Template-directed interference footprinting^{13,14} indicates that NFAT makes strong major-groove contacts only to G1, G2 and T3 (C. Min and G.L.V., manuscript in preparation). Switching of individual A-T base-pairs in the NFAT site to I-C, which drastically alters major-groove functionality but leaves the minor groove unchanged¹⁵, disrupted binding at position 3 but had little or no effect at positions 4–6 (Fig. 3a). Thus, whereas NFAT clearly recognizes the 5'-half site through sequence-specific contacts in the major groove, the protein does not discriminate the 3'-half site through major-groove contacts. These data suggest that NFAT recognizes the 3'-half site either through direct or water-mediated contacts to DNA bases in the minor groove, or by 'indirect readout' of the unique structure of poly(A) DNA. Consistent with either possibility, N3-methylation of any adenine within the 3'-half site interferes with NFAT binding.

To gain insight into the overall architecture of the core NFATc-AP-1-DNA complex, we superimposed the NMR structure of NFATc-DBD on the X-ray structure of the p50-DNA complex, and used the Tyr 29–T3 photocrosslinking observation to position NFATc-DBD relative to AP-1 on ARRE2. A remarkable feature of the resulting model (Fig. 4) is the location of the

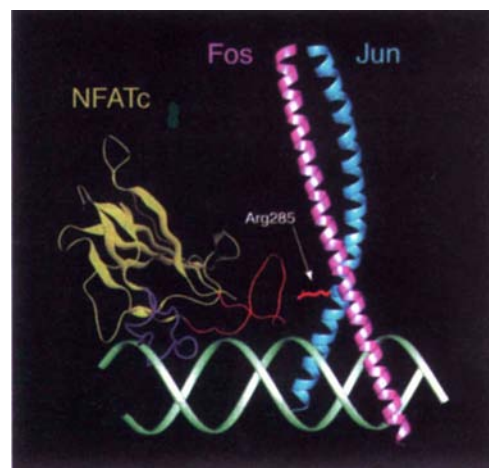


FIG. 4 Model of the core NFATc-AP-1-DNA structure. The solution structure of NFATc-DBD was positioned with respect to the DNA backbone (green) by least-squares superposition of the β -barrel (yellow) with that of a DNA-bound monomer of p50, and with respect to the ARRE2 sequence and AP-1 so as to align Tyr 29 and T3 (as defined by photocrosslinking). The X-ray coordinates of the AP-1 bZip domain²⁹ were used to place the heterodimer (pink, cFos; blue, cJun) over the non-consensus AP-1 site in ARRE2, using positional and orientational information derived from affinity cleaving experiments³⁰. Arg 285 of c-Jun is indicated by an arrow (see text).

NFAT insert region, which ends up positioned directly over the minor groove of the 3'-half site. Although the structure of the insert region is not precisely defined from the NMR data, it is clear that residues of the insert region could readily move within reach of the minor-groove backbone and bases. Furthermore, the two stretches of basic residues at both ends of the insert region (residues 125–128 and 142–146) are reminiscent of minor-groove binding elements commonly used by proteins that recognize DNA by contacting both grooves^{16–18}. Indeed, point mutations within the N-terminal basic stretch (Lys 125, Arg 127 or Asn 128, to Ala) disrupt binding to ARRE2 (L.-J. Sun, B. R. Peterson and G.L.V., manuscript in preparation). We conclude that the NFAT insert region is primarily responsible for sequence-specific recognition of the 3'-half site, and that the insert region of Rel proteins determines whether they bind DNA as monomers (NFAT family) or only as dimers (NF- κ B/c-Rel family).

Whereas NFATc and NFATp exhibit similar specificity for ARRE2, NFAT3 and NFAT4 bind this site relatively weakly at physiological concentrations⁸. Although our discussion has focused on interactions common to all NFAT isoforms, we note that several residues of the DBL are not invariant, and these might help modulate DNA-binding specificity. In random site-selection experiments, NFATp shows a moderate preference for the two nucleotides immediately flanking the 5'-end of the site (positions –1 and –2; C. D. Vaughan, J. Triant and G.L.V., manuscript in preparation), raising the possibility that the characteristic specificity of each NFAT isoform is gained through differential recognition of positions outside the 5'-GGAAAA consensus sequence.

The model of the core NFATc-AP-1-DNA complex also provides insight into the structural basis for cooperativity between NFAT and AP-1. The locus of AP-1 interactions with NFATp maps to the junction of the leucine zipper and basic region¹⁹. Mutation of a single residue (Arg 285) in the c-Jun junction virtually abrogates cooperative interactions between NFATp and AP-1 (ref. 19). In our model, the c-Jun Arg 285 lies closest to the insert region of NFATc-DBD, suggesting that contacts between these loci mediate NFAT-AP-1 cooperativity; this holds true even if the insert structure is folded differently in the DNA-bound state from that in the solution structure of NFATc-DBD. In

agreement with this, point mutations in the NFATc insert region reduce cooperativity between NFAT and AP-1 but have negligible effect on NFATc–DNA interaction (Glu 132 and Thr 138 to Ala; L. J. Sun, B. R. Peterson and G.L.V., manuscript in preparation). Consistent with its critical role in DNA-binding and protein–protein interactions, the insert region is almost identical among the four known NFAT isoforms (Fig. 1c).

The structural basis for the assembly of general transcription complexes on the eukaryotic promoter is becoming clearer^{20,21}. Our results start to show how multiprotein complexes assemble to form the enhancosomes that activate transcription. □

Methods

Overexpression and purification of NFATc. The NFATc DBD used for gel-shift and NMR studies comprises residues 416–591 of human NFATc²⁰, fused to the N-terminal Met1–Lys2 dipeptide. This fragment was cloned into the overexpression plasmid pLM1 with the second codon encoding lysine to improve the level of overexpression. Overexpression was performed in the *E. coli* BL21 DE3 pLysS (Novagen). Transformed cells were grown in LB broth plus 100 mg l⁻¹ ampicillin and 25 mg l⁻¹ chloramphenicol to an A₆₀₀ of 0.6–0.8, and then induced with 225 mg l⁻¹ isopropylthio- β -D-galactoside (IPTG) at 30 °C for 5–6 h. NFATc was purified by cation-exchange chromatography over S-Sepharose (Pharmacia) followed by anion-exchange chromatography over Q-Sepharose (Pharmacia). ¹⁵N- or ¹⁵N/¹³C-labelled NFATc for NMR spectroscopy was prepared as described, except that 1 ml of cells grown in LBAC was used to inoculate 1 litre of M9 minimal medium containing 1 g ¹⁵NH₄Cl and 2 g ¹²C₆- or ¹³C₆-glucose (Cambridge Isotope Labs). ¹⁴N-reverse-labelled NFATc was produced according to ref. 22.

IdU photocrosslinking. IdU-containing oligonucleotides were synthesized by conventional solid-phase chemistry using an IdU phosphoramidite (Peninsula Labs). The IdU-containing strand was ³²P-end-labelled and annealed with its complement. Labelled duplex DNA was mixed with a recombinant fragment of murine NFATp containing the DNA-binding domain⁹, in the presence or absence of a 100-fold molar excess of unlabelled specific or nonspecific competitor oligonucleotide. The mixture was allowed to equilibrate on ice for 1 h, then irradiated on a UV transilluminator (308 nm) for 1 h. The products were loaded directly onto an SDS–polyacrylamide gel (Fig. 3a). The DNA-linked amino acid was identified as described¹¹.

Binding of NFAT to ARRE2 oligonucleotides having A·T base-pairs switched to I·C. ARRE2 oligonucleotides containing individual A·T to I·C substitutions at positions 3 to 6 (Fig. 3a) were synthesized by conventional solid-phase chemistry. The K_d values of NFATc for these oligonucleotides were determined by non-denaturing electrophoretic gel mobility assays. Binding reactions (2 mM HEPES, pH 7.5, 90 mM KCl, 10% glycerol, 0.1% Nonidet-P40, 1 mM DTT, 5 μ g ml⁻¹ dI·dC, 0.1 mg ml⁻¹ BSA) containing various concentrations of NFATc-RHR and 5' ³²P-end-labelled probe (1 nM) were allowed to equilibrate at room temperature for 1 h before loading onto a 6% polyacrylamide gel (0.5 $\times$ TBE). Bands representing free and bound probe were quantified using a Fuji phosphorimager.

NMR spectroscopy. NMR experiments²³ on NFATc-DBD were done with sample concentrations of 1–2 mM in 25 mM deuterated Tris-acetate (Cambridge Isotopes), pH 6.5, 100 mM KCl, 5 mM deuterated DTT. Experiments were done at 27 °C on a Bruker DMX-500 instrument equipped with a triple-resonance probe and triple axis gradients, unless otherwise noted. Assignments for backbone ¹H, ¹³C and ¹⁵N chemical shifts were obtained using a combination of the following 3D experiments: HNCA, HN(CO)CA, HNCO, HN(CA)CO, HCACO, HN(CA)HA. These 3D experiments, in combination with amino-acid assignments determined from comparison of the ¹H-¹⁵N HSQCs of the ¹⁴N reverse-labelled samples²² and from the C α and C β chemical shifts determined from an HC(C)H-TOCSY (400 MHz Varian Unity Plus), allowed the sequential assignment of 156 of the 167 non-proline amino acids. ³J_{H_NHA} coupling constants were obtained from an HNHA experiment and used for backbone dihedral angle constraints. Stereospecific assignments for the methyl groups of 8 of 12 valines and 9 of 19 leucines were established based on the crosspeak pattern in a ¹H-¹³C HSQC (without carbon decoupling) of a 10% ¹³C-labelled sample. H⁺ exchange data were acquired at five time points (1, 2, 12, 36 and 72 h) using a ¹H-¹⁵N HSQC experiment. NOE constraints for structural calculations were obtained from 4 NOESY experiments: 50 ms ¹⁵N separated 3D NOESY-HSQC (750 MHz Varian Unity Plus), 50 ms 2D NOESY in D₂O (750 MHz Varian Unity Plus), 50 ms 2D NOESY in 90% H₂O/10% D₂O, 100 ms ¹³C-separated 3D NOESY-HSQC in D₂O (400 MHz Varian Unity Plus).

Structural calculations. Structural calculations were performed based on 1,087 effective non-hydrogen-bond constraints using the program DIANA²⁴ with three REDAC cycles. Structures with target functions below 10 were further refined by simulated annealing using X-PLOR (version 3.1). 35 structures were

produced with no NOE violations >0.4 Å and no dihedral angle violations >5.0 deg. This family of structures was used for the calculation of statistics, and the ten best structures were chosen for Fig. 1a.

Received 2 October; accepted 12 November 1996.

- Crabtree, G. R. & Clipstone, N. A. *Annu. Rev. Biochem.* **63**, 1045–1083 (1994).
- Müller, C. W., Rey, F. A., Sodeoka, M., Verdine, G. L. & Harrison, S. C. *Nature* **373**, 301–317 (1995).
- Ghosh, G., Van Duyne, G. D., Ghosh, S. & Sigler, P. B. *Nature* **373**, 303–310 (1995).
- Durand, D. B. et al. *Mol. Cell. Biol.* **8**, 1715–1724 (1988).
- Wolfe, S. A. thesis (Harvard Univ., 1996).
- Jain, J., Burgeon, E., Badalian, T. M., Hogan, P. G. & Rao, A. J. *Biol. Chem.* **270**, 4138–4145 (1995).
- Peng, J. W. & Wagner, G. J. *Mag. Res.* **98**, 308–332 (1992).
- Ho, S. N. et al. *J. Biol. Chem.* **270**, 19898–19907 (1995).
- McCaffrey, P. G. et al. *Science* **262**, 750–754 (1993).
- Northrop, J. P. et al. *Nature* **369**, 497–502 (1994).
- Liu, J., Sodeoka, M., Lane, W. S. & Verdine, G. L. *Proc. Natl Acad. Sci. USA* **91**, 908–912 (1994).
- Chen, L. thesis (Harvard Univ., 1994).
- Mascarenas, J. L., Hayashibara, K. C. & Verdine, G. L. *J. Am. Chem. Soc.* **115**, 373–374 (1993).
- Min, C. H., Cushing, T. D. & Verdine, G. L. *J. Am. Chem. Soc.* **118**, 6116–6120 (1996).
- Starr, D. B. & Hawley, D. K. *Cell* **67**, 1231–1240 (1991).
- Pabo, C. O. & Sauer, R. T. *Annu. Rev. Biochem.* **61**, 1053–1095 (1992).
- Blackwell, T. K., Bowerman, B., Priess, J. R. & Weintraub, H. *Science* **266**, 621–628 (1994).
- Feng, J., Johnso, R. C. & Dickerson, R. E. *Science* **263**, 348–355 (1994).
- Peterson, B. R., Sun, L. J. & Verdine, G. L. *Proc. Natl Acad. Sci. USA* **93**, 13671–13676 (1996).
- Nikolov, D. B. et al. *Nature* **377**, 119–128 (1995).
- Tan, S., Hunziker, Y., Sargent, D. F. & Richmond, T. J. *Nature* **381**, 127–134 (1996).
- Shortle, D. J. *Mag. Res. B* **105**, 88–90 (1994).
- NMR of Proteins* (eds Clore, G. M. & Gronenborn, A. M.) (CRC Press, Ann Arbor, 1993).
- Güntert, P., Braun, W. & Wüthrich, K. *J. Mol. Biol.* **217**, 517–530 (1991).
- Hoey, T., Sun, Y.-L., Williamson, K. & Xu, X. *Immunity* **2**, 461–472 (1995).
- Masuda, E. S. et al. *Mol. Cell. Biol.* **15**, 2697–2706 (1995).
- Ghosh, S. et al. *Cell* **62**, 1019–1029 (1990).
- Kieran, M. et al. *Cell* **62**, 1007–1018 (1990).
- Glover, J. N. M. & Harrison, S. C. *Nature* **373**, 257–261 (1994).
- Chen, L. et al. *Curr. Biol.* **5**, 882–889 (1995).

ACKNOWLEDGEMENTS. S.A.W. and P.Z. contributed equally to this work. We thank M. Oliger for assistance in processing relaxation data, S. Park for help with construction of overexpressing strains, and L. J. Sun, B. Peterson, C. Min, D. Erlanson and M. Chytil for discussions and for sharing unpublished data. This work was supported by grants from the NIH and NSF, Howard Hughes Medical Institute, and Hoffmann-LaRoche Institute of Chemistry and Medicine. NMR spectrometers were purchased with funding from the NSF and NIH. S.A.W. was partially supported by a predoctoral fellowship from the NSF; V.D. was supported by a postdoctoral fellowship from EMBO.

CORRESPONDENCE and requests for coordinates and materials should be addressed to G.L.V. (e-mail: verdine@chemistry.harvard.edu).

Structure of the single-stranded-DNA-binding domain of replication protein A bound to DNA

Alexey Bochkarev, Richard A. Pfuetzner, Aled M. Edwards & Lori Frappier

Institute for Molecular Biology and Biotechnology, Cancer Research Group, McMaster University, 1200 Main St. W., Hamilton, Ontario L8N 3Z5, Canada

THE single-stranded-DNA-binding proteins (SSBs) are essential for DNA function in prokaryotic and eukaryotic cells, mitochondria, phages and viruses^{1,2}. The structures of four SSBs have been solved^{3–7}, but the molecular details of the interaction of SSBs with DNA remain speculative. We report here the crystal structure at 2.4 Å resolution of the single-stranded-DNA-binding domain of human replication protein A (RPA) bound to DNA. Replication protein A is a heterotrimeric SSB that is highly conserved in eukaryotes. The largest subunit, RPA70, binds to single-stranded (ss)DNA^{8,9} and mediates interactions with many cellular and viral proteins¹⁰. The DNA-binding domain, which lies in the middle of RPA70, comprises two structurally homologous subdomains oriented in tandem. The ssDNA lies in a channel that extends from one subdomain to the other. The structure of each

TABLE 1 Structure determination and refinement

Data collection*	Native	dC ₈ (I ₂) iso ano		dC ₈ (I ₃) iso ano		dC ₈ (I ₄) iso ano		dC ₈ (I _{2,4}) iso ano	
Resolution (Å)	2.2	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
Total reflections	141,185	37,752	37,755	38,881	38,859	39,663	39,651	45,730	45,701
Unique reflections	13,045	5,311	9,366	5,397	9,571	5,374	9,569	5,431	9,652
Completeness (%)†	95.0/87.8	95.2/90.2	93.7/84.2	96.7/93.2	95.7/90.1	96.3/94.7	95.7/92.3	97.3/97.0	96.5/93.7
$I/\sigma(I)$ †	28.0/7.8	23.2/11.8	18.3/9.1	24.7/13.4	19.0/10.3	27.4/17.2	23.2/13.9	29.4/16.7	24.7/13.2
R_{sym}	6.3	6.1	5.3	6.6	5.9	5.5	4.9	5.8	4.8
R_{iso}			11.9		8.9		9.7		15.2
R_{ano}			3.8		3.1		3.0		3.4
Phasing (30.0–30.0 Å)									
Number of heavy atoms		1	1	1	1	1	1	2	2
Cullis R factor		63.9		60.0		61.5		64.7	
Kraut R factor		10.1	11.6	7.3	8.4	8.4	11.3	13.4	18.8
Phasing power		1.22	1.42	1.33	1.10	1.17	1.22	1.13	2.01
Mean FOM after MIRAS phasing	0.764								
Refinement									
Resolution (Å)	6.0–2.4								
R factor	21.2								
R -free	33.0								
Reflections (with $F > 2\sigma(F)$)	9,460								
Completeness (%)†	96.5/93.4								
Total number of non-hydrogen atoms	2,122								
Water molecules	91								
R.m.s. deviations (protein/DNA)									
Bond length (Å)	0.016/0.033								
Bond angles (°)	1.60/3.45								

$R_{\text{iso}} = \sum |F_{\text{ph}}^2 - F_{\text{p}}^2| / \sum (F_{\text{ph}}^2 + F_{\text{p}}^2)$; $R_{\text{ano}} = \sum |F(+)^2 - F(-)^2| / \sum (F(+)^2 + F(-)^2)$; $R_{\text{sym}} = \sum (|I_{\text{obs}} - I_{\text{avg}}|) / \sum I$ (summation over all reflections); R factor = $\sum ||F_{\text{obs}}| - |F_{\text{calc}}|| / \sum |F_{\text{obs}}|$; $R_{\text{cullis}} = \sum (||F_{\text{ph}} \pm F_{\text{p}}| - |F_{\text{h,calc}}||) / \sum (|F_{\text{ph}} \pm F_{\text{p}}|)$ (centric reflections only); $R_{\text{kraut}} = \sum |F_{\text{ph,obs}} - F_{\text{ph,calc}}| / |F_{\text{ph,obs}}|$ (acentric reflections only); phasing power is r.m.s. $F_{\text{h}}/\text{r.m.s. residual lack of closure}$.

* Data processing using only reflections $I > \sigma I$; unit cell parameters: $a = 34.3$, $b = 78.0$, $c = 95.4$ Å; space group $P2_12_12_1$.

† Overall/outer shell

RPA70 subdomain is similar to those of the bacteriophage SSBs, indicating that the mechanism of ssDNA-binding is conserved.

The DNA-binding domain of the largest subunit of the RPA heterotrimer was identified using a combination of limited proteolysis and deletion analysis^{8,9,11}. A complex of this domain (RPA70_{181–422}) with octadeoxycytosine was crystallized¹¹ and the structure solved using isomorphous and anomalous scattering from complexes with four different iodinated oligonucleotides (Table 1). Electron density was observed for the entire DNA and the entire protein, except for the first six residues (four of which were derived from the fusion protein) and the last two residues (Fig. 1).

RPA70_{181–422} is composed of two structurally similar domains, each of which binds ssDNA (Fig. 2). The r.m.s. deviation between the two domains, which were predicted to be comparable on the basis of sequence similarity¹², is 1.495 Å for 74 residues (Fig. 3). The first domain (domain A) extends from amino-acid residue I198 to D291, and the second (domain B) from I305 to N402. Domain A has a disulphide bridge between C200 and C289. The core of each domain is composed of an OB-fold, which consists of a '5-stranded β -sheet that is coiled to form a closed β -barrel' and an α -helix that connects the third and fourth strands¹³ (Figs 2 and 3).

The DNA traverses the surface of β -strands $\beta 1$ to $\beta 3$. The $\beta 1'$ and $\beta 2$ strands of each sheet extend like fingers to form a β -hairpin that wraps around the DNA and creates a DNA channel that is ~ 17 Å across (Fig. 2b). The DNA-binding channels in the two domains are oriented in the same direction and consequently the DNA extends from one domain to the other in approximately a straight line. In the co-crystal, three nucleotides are contained

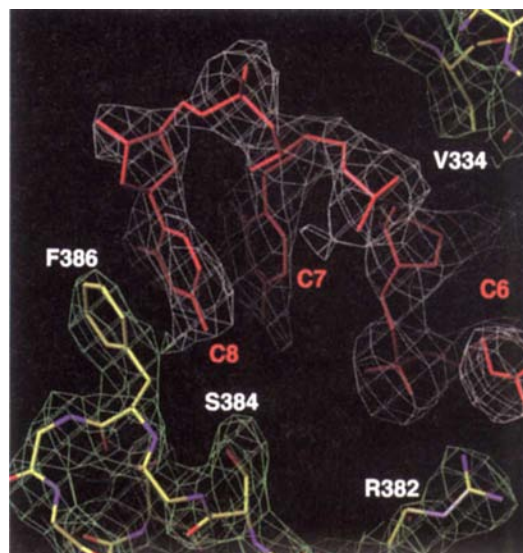


FIG. 1 Region of electron density in the solvent-flattened 3.0 Å experimental map that shows the interaction of RPA with the DNA. The protein density is shown in green and the DNA density in white. Superimposed on the map is the refined model of the protein and the DNA. Contours drawn at the 1 σ level.

within each domain and the space between the domains is bridged by two nucleotides. The two domains are connected by a 15-amino-acid linker on the opposite side of the protein with respect to the DNA (Fig. 2). The C-terminal portion of the fragment that was crystallized forms an α -helix (residues 407–420). This helix plays no role in DNA binding¹¹, but its inclusion in our expression construct improved the solubility properties of the resulting protein.

The long axis of the DNA follows a relatively straight path from one domain to the other. The sugar–phosphate backbone forms a ridge on one face of the DNA. The ridge is slightly S-shaped with kinks between the third and fourth, and fifth and sixth bases. All eight of the sugar–base bonds are pointed in the direction of the bulk of the protein, confined to a 120° arc around the axis of the sugar–phosphate backbone (Fig. 2*b*). The planes of the DNA bases are arranged in three sets of pairwise stacking interactions: base 2 stacks with base 3, base 4 with base 5, and base 7 with base 8;

the distances between each set of aromatic rings range from 3.5 to 4.2 Å. The DNA does not have any standard configuration, but the interphosphate distance of 5.4–6.3 Å in between the third and fourth, the fourth and fifth, and the fifth and sixth phosphates is more consistent with a 3'-endo sugar pucker observed in A-form DNA. A complete assignment of the sugar conformations must await further refinement at higher resolution.

The three nucleotides (dC₁₋₃ and dC₆₋₈) that reside within each of the two DNA-binding channels have a very similar structure (Fig. 2*a*). The position of these nucleotides is maintained by an extensive network of hydrogen-bonding with RPA (Fig. 4), as well as stacking interactions between phenylalanine residues and the bases (Fig. 5). However, although the structure of the trinucleotide bound by each domain is almost identical, and the overall fold of the two domains is very similar, the protein/DNA contacts for each domain are considerably different (Fig. 4).

Domain A makes much more extensive contact with the DNA.

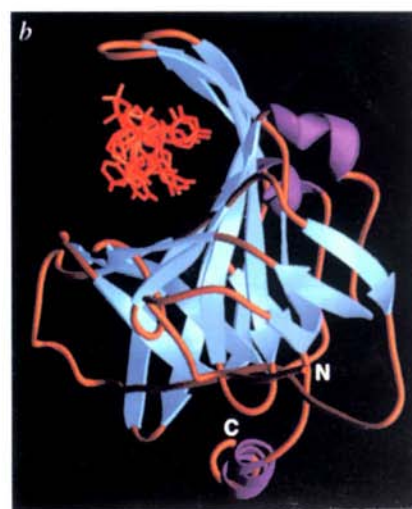
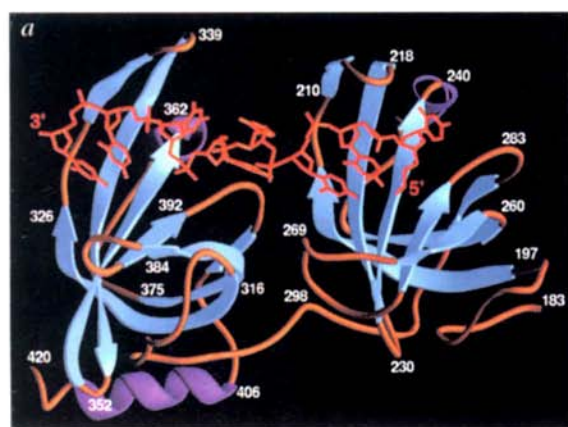


FIG. 2 The RPA/DNA complex. *a*, Ribbon diagram with RPA coloured in blue (β -strands), purple (α -helices) and gold (loops) and the DNA coloured in red. Amino-acid numbers are shown in white. *b*, View down the axis of the DNA-binding channel.

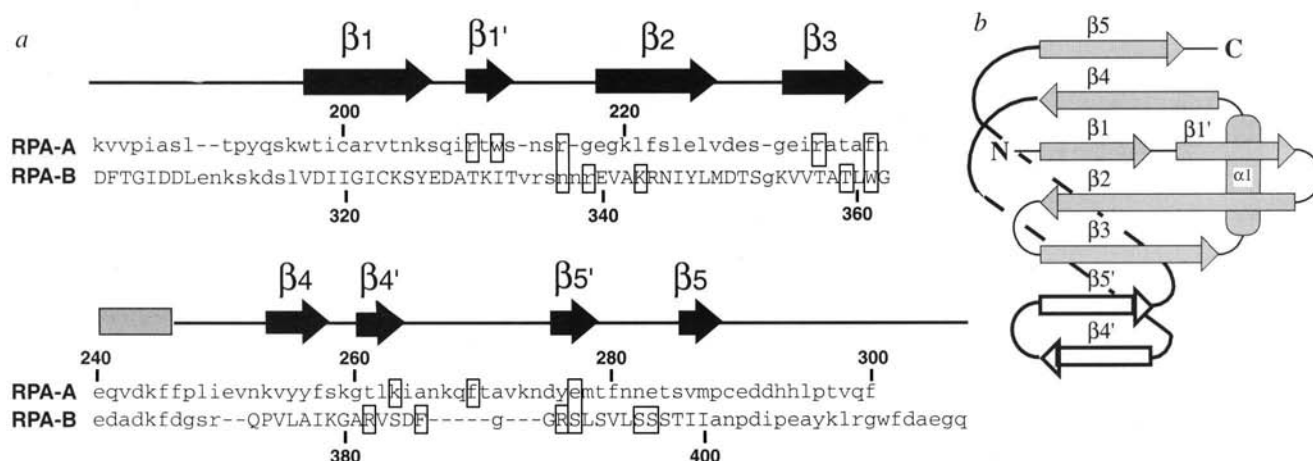


FIG. 3 The RPA subdomains. *a*, A sequence alignment of the RPA domains and other SSBs. The boxed residues are those that interact with the DNA (Fig. 4). The residues that are capitalized in RPA-B indicate those that were used to do an α -carbon structural alignment using the lsq_explicit feature in the molecular graphics program O (ref. 20) (r.m.s. deviation, 1.495 Å for 74

atoms). *b*, Topology of the RPA domains. The β -strands are indicated by arrows and the α -helix by an oval. The hatched β -strands correspond to those that comprise the OB fold. The DNA-binding β -hairpin comprises β_1' and β_2 . The extended loop that contains β_4' and β_5' is referred to as the L₄₅ loop in the nomenclature of the OB fold¹³ and is indicated in bold.

In this domain, two phenylalanine residues (F238 and F269) stack with the DNA bases. F238 stacks with dC₁ (3.7 Å between the planes of the rings). F269 stacks with dC₃ (3.5 Å between the planes of the rings), which in turn stacks with dC₂, rather like a set of three dominos on end (4.2 Å between the planes of the dC₂ and dC₃ rings; Fig. 5). R210 makes an interesting interaction with the DNA: the R210 side chain hydrogen-bonds with the phosphate from the third base (distance of 2.83 Å). This hydrogen-bond interaction orients the plane of the R210 guanidino group in a position to stack with the plane of the ring of the dC₄, which in turn stacks with the ring of the dC₅.

The hydrogen-bonding amino acids are located in both the β -hairpin and in the β 1– β 3 sheet. There are a total of six hydrogen bonds to the phosphates in dC_{1–3} (Fig. 4). The most remarkable set of interactions are hydrogen bonds between the protein and the DNA bases. In particular, two side chains in domain A, E277 and R234, hydrogen-bond with the cytosine of dC₂ and dC₃. The side chain of E277 forks towards the N4 of each of dC₂ and dC₃, and the forked side chain of R234 binds to the O2 carbonyl from both dC₂ and dC₃ (Figs 4, 5). These residues apparently supply the hydrogen-bond donors and acceptors normally contributed by the complementary Watson–Crick bases.

Domain B makes fewer contacts with the DNA compared with domain A. The aromatic residues that make important stacking contacts in domain A (F238 and F269) are conserved in domain B (W361 and F386 respectively). However, whereas in domain A the rings of F238 and dC₁ are parallel, in domain B the planar rings of W361 and the analogous base dC₆ are at a 55° angle with respect to each other. F386 in domain B, by analogy with F269 in domain A, partakes in the domino-stacking interaction with dC₈ and dC₇. The middle base of the domino triad, dC₈, is slightly displaced from the row.

Domain B makes three hydrogen bonds to the phosphates of dC_{6–8} and four hydrogen bonds to the bases in dC_{6–8} (Fig. 4). The pattern of hydrogen bonds to the bases differs between domains A and B. For example, in domain A, R234 forms hydrogen bonds with each of the O2 carbonyls of dC₂ and dC₃; in domain B, the equivalent interaction (K343) is only to dC₇ and no residue plays the role of E277, which binds to each of the N4 residues of these bases.

We believe that some of the differences between the interac-

tions of domains A and B with the DNA result from crystal-packing interactions. The base of dC₈ makes no direct hydrogen bonds with domain B. In fact, the phosphate of dC₈ is hydrogen-bonded to a lysine residue derived from another RPA molecule. This interaction appears to pull the dC₈ base out of stacking register, which has two consequences. First, the domino relationship between F386, dC₈ and dC₇ is disrupted, and second, the dC₈ base may be pulled away from its normal hydrogen-bonding partners.

Domain B makes a total of six hydrogen bonds to the two nucleotides between domains A and B (dC₄ and dC₅). Specifically, the O2 and N4 positions of dC₄, the phosphate of dC₅, as well as the O2 and N4 from the dC₅ base are involved in hydrogen bonds (Fig. 4). The interaction of domain B with two residues upstream suggests by analogy that domain A may be capable of similar interactions. The presumed interactions of domain A with upstream bases could not be seen in our crystal structure because the dC₈ oligonucleotide does not extend upstream beyond the domain A DNA-binding channel.

The phosphate backbone is solvated by a string of water molecules. Several of these waters mediate protein/DNA hydrogen-bond interactions. For example, the 5' phosphate interacts with S279 through a water molecule. The analogous phosphate in domain B, the phosphate from dC₆, is bound to two water molecules that are in turn hydrogen-bonded with amino acids R382 and S392. The dC₇ phosphate interacts with S336 through a water molecule and the O2 carbonyl of the eighth base interacts with N345 through a water.

Despite the fact that both the DNA and the protein backbones are virtually identical in structure, domains A and B make considerably different contacts with the DNA. For example, K343 and R234 make hydrogen bonds to equivalent carbonyls in the DNA, but K343 and R234 do not occupy topologically equivalent positions in domains A and B. There is also no analogue for E277 in domain B. These differences suggest that the binding of the RPA domains to DNA has some plasticity. There may be many residues in RPA that are capable of interacting with DNA, of which we have captured only a subset in our crystal structure. Perhaps alternative sets of protein/DNA interactions are used to accommodate binding to thymine or the purine bases. There also may be considerable changes in the structure

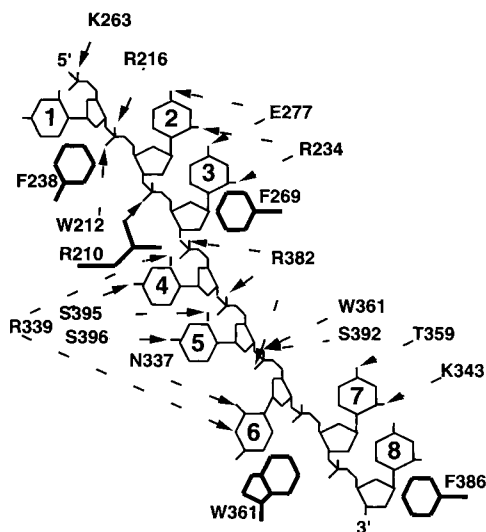


FIG. 4 Diagram showing the contacts of RPA with dC₈. The bases are numbered 1–8 starting from the 5' end, and the numbers have been placed in the corresponding bases. Hydrogen bonds between the amino-acid side chains and the DNA are marked by dashed lines. The aromatic residues that stack with the bases are marked in bold.

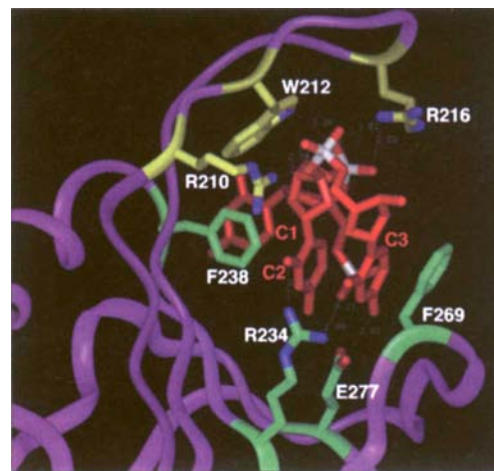
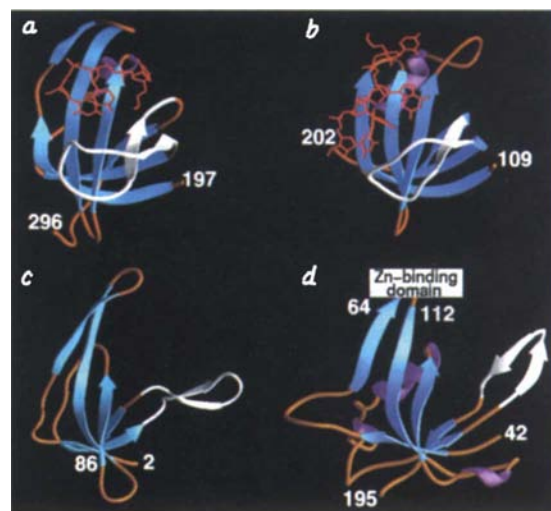


FIG. 5 Model showing the interaction of the RPA domain A with DNA (generated using INSIGHT II; Biosym, San Diego, California). DNA sugars and bases are coloured in red and phosphorus atoms in white. The protein backbone is shown in purple. Amino-acid residues that form hydrogen bonds—or stack with—the DNA bases are shown in green, and those that interact with the DNA phosphates are yellow. Within the side chains, oxygen atoms are red and nitrogen atoms blue. Hydrogen bonds are shown as dashed white lines and the hydrogen-bond distances are indicated.

FIG. 6 Comparison of RPA-A (a), yeast aspartyl-tRNA synthetase (b), the gene V protein from bacteriophage f1 (c) and the gp32 protein from bacteriophage T4 (d). Ribbon representations of the four structures are shown. Structures are coloured as in Fig. 2, except that the region that corresponds to the L_{45} loop in the OB fold is white. In a and b, the nucleic acid is shown in red. Atomic coordinates for the structures in b, c and d were obtained from the Protein Data Bank (entries 1asy, 1bgh and 1gpc, respectively).



when bound to longer DNA. An accounting of the protein/DNA contacts with DNA molecules of different size and sequence awaits further structural and biochemical analysis.

Our crystal structure reveals that RPA_{70,181-422} can form a stable complex with oligonucleotides as short as an 8-mer, which is consistent with previous results²³. Our DNA-binding studies did not reveal any difference between the binding affinity of RPA₁₈₁₋₄₂₂ for an 8-mer or an 18-mer¹¹, which suggests that an octanucleotide is very close to the optimum length. But although the crystal structure suggests that RPA₁₈₁₋₄₂₂ interacts with 8–10 nucleotides, under certain solution conditions the length of DNA that is occluded by heterotrimeric RPA can be as long as 30 or 90 nucleotides^{14,15,24}. These larger binding sites may result from altered conformations of RPA that might be adopted under different solutions conditions: our structural information does not address this possibility. Alternatively, other regions of RPA70 or other RPA subunits¹² may increase the effective size of the DNA interaction site by either making additional protein–DNA interactions or by sterically limiting the accessibility of other RPA molecules for the DNA.

The RPA subdomains differ in two ways from the described five- β -strand OB fold¹³. First, in both RPA domains, the first β -strand of the OB fold is broken into two strands that are separated by a short linker sequence. Second, in the described OB fold¹³, a short loop, called the L_{45} loop, connects the fourth and fifth β -strands. This region of the OB fold is implicated in substrate binding. In RPA, the ' L_{45} loop', which connects β -strands 4 and 5 in the context of RPA, is quite large, contains two small β -strands (β_4' and β_5'), and folds back so that β_5' runs parallel with the third β -strand of the RPA domain. Of the many proteins in the protein structure database that have an OB fold, RPA most resembles the anticodon-binding domain of yeast aspartyl tRNA synthetase bound to RNA¹⁶ (Fig. 6a, b). The r.m.s. deviation between RPA-B and the aspartyl-tRNA synthetase is 1.57 Å over 84 of the 119 residues. The aspartyl-tRNA synthetase and RPA use common mechanisms to bind the single-stranded nucleic acid, including aromatic stacking interactions and hydrogen bonds to the bases and the phosphates. In addition, the residues that interact with the nucleic acids are located in common positions, in the β -hairpin and the L_{45} loop.

RPA and the aspartyl-tRNA synthetase represent the only OB-fold-containing proteins whose structures have been solved bound to an oligonucleotide. However, this fold has also been described in SSBs whose structures have been solved in the absence of DNA, including the bacteriophage Pf3 protein and the gene V protein from f1 phage³⁻⁵. These two phage SSBs are very similar in structure and function⁶ and their topologies are virtually identical

with those of RPA and the tRNA synthetase. The similarities extend even to the topological location of the residues that are important for DNA binding. In all four proteins, the critical residues for DNA binding, as determined by either site-directed mutagenesis or from the analysis of the crystal structure, are located along the β -hairpin and within the L_{45} loop.

Although the three-dimensional positions of the β -hairpin and the β -sheet are conserved among the gene-V, Pf3, RPA and tRNA synthetase structures (for example the r.m.s. deviation in this region between the gene V protein and RPA is 1.86 Å for 38 residues), there are dramatic differences in the relative positions of the L_{45} loops (Fig. 6a, b, c; L_{45} loops are shown in white in Fig. 6). In the gene V and Pf3 proteins, which are dimers, the L_{45} loops from one monomer extends away from the OB fold to interact with the DNA-binding channel in the other monomer^{3,6}; the tip of the L_{45} loop constitutes the dimerization interface. These dimeric proteins are predicted to bind two strands of DNA travelling in opposite directions. In RPA and the aspartyl-tRNA synthetase, the L_{45} loop does not extend away from the molecule but rather doubles back to form part of the DNA-binding β -sheet (Fig. 6).

The structural similarities and differences between the f1 phage proteins and RPA/aspartyl-tRNA synthetase suggests that there may be two modes of DNA binding by this topological class of structures. The 'RPA/aspartyl-tRNA synthetase' mode, in which the OB fold is active as a monomer, might favour binding to a single DNA molecule, whereas the 'f1 phage' dimeric mode may favour binding to two strands in opposite directions. Alternatively, the single-stranded nucleic acid OB fold may have two conformations. The structures of the gene V and Pf3 phage proteins, which were solved in the absence of DNA, might represent the non-DNA-bound, or open, form of the SSB. The RPA structure might represent the DNA-bound, or closed, form of the SSB. In this model, a large conformational change in the position of the L_{45} loop would be predicted to accompany DNA binding. We favour the open/close model because it offers (1) an explanation for the virtually identical disposition of critical amino acids in the topological map of the phage and human SSBs, despite the differences in the three-dimensional structures, and (2) a mechanism by which the single-stranded DNA might load onto SSBs.

The conservation of structure and function among SSBs is supported by the structure of the T4 phage gp32 protein, which was solved in the presence of DNA⁷. The topology and three-dimensional structure of the gp32 N-terminal domain is very similar to that of the other proteins, except that gp32 protein has a large zinc-binding subdomain at the top of the β -hairpin (indicated in Fig. 6d). In the gp32 co-crystal, the electron density for the DNA was weak, implying that the DNA was either

disordered or had low occupancy in the crystal. The gp32 L₄₅ loop is in the open conformation. We suggest that the structure of the gp32 protein bound to DNA represents that of a binding intermediate, that of DNA binding weakly to the open conformation. In the co-crystal, the DNA may have been disordered because the gp32 L₄₅ loop was unable to close to stabilize the binding, perhaps because of the geometry of crystal packing, or alternatively because of unfavourable solution conditions.

Finally, it is likely that the interaction of RPA with single-stranded DNA has polarity. The OB folds of both RPA and the aspartyl-tRNA synthetase are oriented identically with respect to the 5' and 3' ends of the DNA. In terms of the DNA replication reaction, if RPA was bound to the template strand in this orientation, domain B of RPA would be proximal to the advancing DNA polymerase and the N-terminal region of RPA70, which is implicated in regulatory protein-protein interactions¹⁰, would be distal from the replicating polymerase. □

Methods

Crystallization and data collection. RPA₁₈₁₋₄₂₂ was cloned into the T7 polymerase expression vector, pET15b (Novagen), and overexpressed in the bacterial strain BL21(DE3) pLysS as a fusion to an N-terminal six-histidine tag¹⁷. The protein was purified by a combination of nickel-chelate affinity chromatography and, after removal of the histidine tag by digestion with thrombin, cation-exchange chromatography¹⁸. Four amino acids, GSHM, persist from the fusion protein and extend from the N terminus of RPA. The protein was concentrated using a Centricon ultrafiltration membrane and stored frozen at 20 mg ml⁻¹ in a solution containing 5 mM HEPES, 1 mM DTT, 200 mM NaCl, pH 6.9. The octadeoxycytosine, dC₈ (Dalton Chemical) was purified by reverse-phase HPLC, phosphorylated with polynucleotide kinase, extracted with phenol/chloroform and precipitated with isopropanol in the presence of 0.1 M MgCl₂. Iodinated DNA was treated in the same way. Complexes of RPA₁₈₁₋₄₂₂ (4–8 mg ml⁻¹) and DNA (1.5-fold molar excess) were formed at 4 °C for 20 min. Crystals of the RPA/DNA complex were grown at 4 °C by vapour diffusion in sitting drops against a reservoir solution containing 24–27% PEG2000 MME, 20 mM NaCl, 20 mM MgCl₂, 5 mM DTT, and buffered with 100 mM bis-Tris, pH 6.7, 100 mM MES, pH 6.8, or 100 mM HEPES, pH 6.9 or pH 7.0. The sitting drops were formed by mixing 5 µl of the RPA/DNA complex and 5 µl of reservoir solution. Crystals formed within days and in a few weeks typically grew to 0.2 mm × 0.2 mm × 0.8 mm. The crystals (space group *P*2₁2₁2₁; *a* = 34.3 Å, *b* = 78.0 Å, *c* = 95.4 Å) contained one RPA/DNA complex in the asymmetric unit and diffracted beyond 2.2 Å. For low-temperature data collection, crystals were cryoprotected in reservoir solution augmented to 35% PWG2000 MME before being mounted in a loop made of a fibre of dental floss and flash-frozen in a stream of N₂ gas. Each data set was collected from a single crystal maintained at –175 °C on a Rigaku RAXIS-IIc image plate area detector with a rotating anode X-ray source (Table 1). The diffraction data were integrated with DENZO and the intensities scaled with SCALEPACK (Z. Otwinowski). Four heavy-atom derivatives, three different mono-iodinated derivatives, and a di-iodinated derivative were used in the structure determination. Mono-iodinated derivatives were prepared by co-crystallizing RPA with dC₈ iodinated at one of positions 2, 3 or 4. The di-iodinated derivative was prepared by co-crystallizing RPA with dC₈ iodinated on positions 2 and 4. A data set was also collected from a selenium-containing crystal. Selenium was incorporated into RPA by expressing and

purifying RPA from B834 (Novagen), a *met* auxotrophic strain of *E. coli*, induced in the presence of selenomethionine. Using data from frozen native and derivative crystals, the iodine positions were determined from isomorphous difference Patterson and anomalous difference Patterson maps. The phases from the iodinated derivatives were used to locate the positions of four selenium atoms by difference Fourier analysis. The positions of these selenium atoms were used to guide the map interpretation. After maximum-likelihood refinement of heavy-atom, scaling and lack-of-closure parameters, the mean figure of merit was 0.769 for the resulting MIRAS phase set. A final solvent-flattened MIRAS map at 3 Å resolution was then obtained by 16 averaging, flattening and phase-combination cycles. All heavy-atom refinement, phasing, solvent-flattening and map calculations were carried out using the PHASES program package¹⁹. The electron density map was readily interpretable at 3.0 Å and showed continuous electron density for the DNA and most of the protein. A model was built using the program O (ref. 20) and then refined against the 2.4-Å data set by a combination of simulated annealing (XPLOR)²¹, conventional refinement and manual fitting. The protein parameters of Engh and Huber²² and modified DNA parameters (S. K. Burley, personal communication) were used for refinement. The refined model has an r.m.s. deviation from ideality of 0.016 Å and 1.6 degrees for bond distances and angles, respectively. Six residues from the N terminus (two derived from RPA and four from the fusion protein) had no density and were not included in the final model. Two residues from the C terminus were also not visible in the structure. The final *R* factor is 0.212 for all 9,460 reflections between 6 and 2.4 Å; the Ramachandran plot shows two violations (amino acids 239 and 259), with over 85% of all residues in most favoured regions.

Received 11 September; accepted 11 November 1996.

- Kornberg, A. & Baker, T. *DNA Replication* 2nd edn (Freeman, New York, 1992).
- Karpel, R. L. in *The Biology of Non-Specific DNA-Protein Interactions* (ed. Revzin, A.) 103–126 (CRC, Boca Raton, Florida, 1990).
- Skinner, M. M. et al. *Proc. Natl Acad. Sci. USA* **91**, 2071–2075 (1994).
- Folkers, P. J. M., Nilges, M., Folmer, R. H. A., Konings, R. N. H. & Hilbers, C. W. J. *Mol. Biol.* **236**, 229–246 (1994).
- Tucker, P. A. et al. *EMBO J.* **13**, 2994–3002 (1994).
- Folmer, R. H. A., Nilges, M., Konings, R. N. H. & Hilbers, C. W. *EMBO J.* **14**, 4132–4142 (1995).
- Shamoo, Y., Friedman, A. M., Parsons, M. R., Konigsberg, W. H. & Steitz, T. A. *Nature* **376**, 362–366 (1995).
- Gomes, X. V. & Wold, M. S. *J. Biol. Chem.* **270**, 4534–4543 (1995).
- Gomes, X. V., Henrickson, L. A. & Wold, M. S. *Biochemistry* **35**, 5586–5595 (1996).
- Wold, M. *Annu. Rev. Biochem.* **66**, 61–91 (1997).
- Phuetzner, R. A., Bochkarev, A., Edwards, A. M. & Frappier, L. *J. Biol. Chem.* (in the press).
- Philipova, D. et al. *Genes Dev.* **10**, 2222–2233 (1996).
- Murzin, A. G. *EMBO J.* **12**, 861–867 (1993).
- Kim, C., Snyder, R. O. & Wold, M. S. *Mol. Cell. Biol.* **12**, 3050–3059 (1992).
- Alani, E., Thresher, R., Griffith, J. D. & Kolodner, R. D. *J. Mol. Biol.* **227**, 54–71 (1992).
- Ruff, M. et al. *Science* **252**, 1682–1689 (1991).
- Studier, F. W., Rosenberg, A. H., Dunn, J. J. & Dubendorff, J. W. *Meth. Enzymol.* **185**, 60–89 (1990).
- Barwell, J. A. et al. *J. Biol. Chem.* **270**, 20556–20559 (1995).
- Furey, W. & Swaminathan, S. *Meth. Enzymol.* (in the press).
- Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. *Acta Crystallogr. A* **47**, 110–119 (1991).
- Brünger, A. T., Kuriyan, J. & Karplus, M. *Science* **235**, 458–460 (1987).
- Engh, R. A. & Huber, R. *Acta Crystallogr. A* **47**, 392–400 (1991).
- Blackwell, L. J. & Borowiec, J. A. *Mol. Cell. Biol.* **14**, 3993–4001 (1994).
- Blackwell, L. J. & Borowiec, J. A. & Mastrangelo, I. A. *Mol. Cell. Biol.* **16**, 4798–4807 (1996).

ACKNOWLEDGEMENTS. This work was supported by grants to L.F. and A.M.E. from the National Cancer Institute and the MRC of Canada (NCIC and MRC). A.B. was supported by a fellowship from the NCIC; L.F. is a research scientist of the NCIC; A.M.E. is a research scholar of the MRC. We thank M. Wold, G. Borgstahl and C. Arrowsmith for comments on the manuscript. Atomic coordinates have been deposited in the Protein Data Bank (JNMC).

CORRESPONDENCE and requests for materials should be addressed to A.M.E. (e-mail: edwardsa@fhs.mcmaster.ca) or L.F. (e-mail: frappier@fhs.mcmaster.ca).

Guide to authors of contributions to *Nature*

Nature is a weekly international journal covering all the sciences. It is intended for an interdisciplinary readership, so all manuscripts should be written clearly and simply. Contributors should pay particular attention to the fact that English is not the first language of many readers. Space is limited and competition severe, so brevity is highly valued. The length limits below are specified in terms of the number of pages taken up in *Nature*. An uninterrupted page of text contains about 1,300 words, so a typical Letter to *Nature*, occupying 2½ pages, can contain about 1,500 words of text and four modest display items (figures and/or tables) with brief legends. There are no formal limits on the number of display items, but extra display material must be accompanied by commensurate reductions in the length of the text. Length limits can, however, be varied upwards or downwards at the editors' discretion.

Nature publishes the following types of contribution:

Review Articles survey recent developments in a topical area of scientific research. They should not generally occupy more than 6 pages of *Nature*.

Progress articles are relatively informal reviews of particularly topical and fast-moving fields. They do not exceed 4 pages of *Nature*.

Contributors wishing to submit a Review Article or Progress should first send a one-page synopsis to the Reviews Coordinator, from whom guidelines are available on request.

Articles are reports of original scientific research whose conclusions represent a substantial advance in understanding of an important problem and are of broad general interest. They do not generally exceed 5 pages of *Nature* and have no more than 50 references. They have a 'heading' of up to 80 words, often rewritten by *Nature's* editors, which advertises the content of the paper in general terms and should not contain numbers, abbreviations or measurements unless central to the message of the paper. The introduction and summary should be contained in the first three or four paragraphs of the main text. Layout of methods, figures and legends is summarized below.

Letters to *Nature* are short reports of original research focused on an outstanding finding whose implications are of interest to nonspecialists. They should not generally exceed 2½ pages of *Nature*. They have a bold first paragraph of not more than 180 words summarizing the background, rationale, conclusions and main implications of the work, aimed particularly at nonspecialist readers. Letters should have no more than 30 references. Layout of methods, figures and legends is summarized below.

Preparation of Articles and Letters

All manuscripts should be written in English, typed double-spaced on one side of the paper only. Manuscripts submitted in electronic form cannot be considered.

Manuscripts should be accompanied by a brief covering letter from the corresponding author, containing a full postal address, telephone and fax numbers, and e-mail address.

Five copies of manuscripts and original figures are needed, together with two copies of the covering letter. Five copies of manuscripts by any of the authors that have been submitted or are in press elsewhere should also be included, clearly marked as such; failure to do so may prejudice consideration of the manuscript.

Unless otherwise instructed by the editor handling a manuscript, when sending revised or resubmitted manuscripts five copies are required, each accompanied by a copy of the authors' response to referees' and editors' comments on the earlier version.

Titles should not exceed three lines of 31 characters each (including spaces), and should avoid numerical values, acronyms, abbreviations and punctuation wherever possible.

References should be numbered sequentially as they appear in the text, tables and figure legends. Only papers that have been published or submitted to a journal should be in the numbered list: manuscripts in preparation should be mentioned in the text with a list of authors. Reference lists should not contain textual material, grant details or acknowledgements.

Acknowledgements must be brief and appear after the reference list. *Nature* does not publish grant or contribution numbers.

Figures should not be larger than 22 by 28 cm unless unavoidable. They should not contain more than one panel unless the assembled parts are logically connected. Each copy should be marked with the corresponding author's name and, when known, the manuscript reference number. One photocopy of the original figures should be provided. All original figures are returned when *Nature* cannot offer to

publish a manuscript, but one copy of the manuscript and a photocopy of the figures are kept in confidential files for three months and then destroyed. Detailed instructions for figure preparation are available from the production department, including details of acceptable digital formats.

Colour figures are welcome but a contribution towards the cost of reproduction is requested. Inability to pay this charge will not prevent *Nature* publishing figures where colour is essential.

Tables. Symbols and abbreviations in a table should be defined immediately below the table, followed by essential descriptive material, all in double-spaced text. Tables should each be presented on a separate sheet of paper.

Methods. Brief details of materials and methods can be given in the text. Otherwise, they should be described at the end of the text in a section entitled "Methods", subdivided by bold headings referring to the methods used. This section should not exceed 800 words and ideally should be shorter. Additional material can be deposited on *Nature's* Web site. In contributions with methods sections, figure legends should not exceed 150 words each, and should be confined to a description of the figure and the symbols used.

Materials. As a condition of publication, authors are required to make materials and methods used freely available to academic researchers for their own use. Supporting datasets must be made available at the time of publication from the authors directly, by posting on *Nature's* Web site, by deposition in the appropriate public database or by distribution on the Internet. In the case of X-ray crystallographic coordinates only, public access may be delayed for up to one year after publication.

Articles and Letters should both contain a statement at the end of the text: "Correspondence and requests for materials to xxxx", with an e-mail address if desired, followed by the accession numbers or Web site address for supporting datasets. The designated author need not necessarily be the senior author, but *Nature* will expect the named author to handle matters arising from published contributions.

Manuscripts can be submitted to the Editor at *Nature*, Porter's South, Crinan St, London N1 9XW or at *Nature*, 968 National Press Building, Washington, DC 20045-1938. Proofs should be returned by express mail to London. Fax numbers and e-mail addresses for all editorial offices are on *Nature's* masthead.

Other contributions

Nature publishes informal contributed material as follows:

Correspondence contributions are short comments on topical issues, anecdotal material or reactions of readers to nontechnical material published in *Nature* (mainly News, Briefings, Opinion, Book Review or Commentary articles). All are unsolicited; submissions by e-mail to corres@nature.com are preferred (accents on names and places should be clearly indicated).

Commentary articles are informal, editorial comments on topical issues of public interest that have some bearing on scientific research. They are often controversial. Unsolicited manuscripts are considered, but enquiries can be made to the Commentary Editor by e-mail (nature@nature.com) or fax as a one-page synopsis or informal letter before formal submission.

News and Views articles inform nonspecialist readers about new scientific advances, either as reported in published papers or in the form of meeting reports. Most are commissioned, but proposals can be made to the News and Views Editor in advance of publication of the paper or of the meeting concerned. Authors are not allowed to discuss their own work or work from elsewhere in their institution. Detailed guidelines are available on request.

Scientific Correspondence is a predominantly peer-reviewed section of the journal specializing in very brief reports of general interest and discussion of topical issues. Priority is given to contributions of fewer than 500 words. Detailed guidelines are available on request. All contributions are unsolicited, and should be submitted by mail, not e-mail. Copies of contributions submitted by fax should be sent by mail with original figures and a cover letter noting that the manuscript has been faxed.

Book Reviews are all commissioned by the Book Review Editor. Unsolicited contributions are not considered.

Nature's editors will request final copy for Commentary, News and Views, Scientific Correspondence and Book Reviews on disk or via e-mail, as appropriate.

Making 'antisense' of amplification

This edition covers items for DNA amplification, such as an RT-PCR system, a sequence detection system, a DNA decontamination kit, DNA isolation and concentration tools, and items for maintaining clean PCR work areas.

STRATAGENE has introduced the Hot Top assembly for the RoboCycler 40 **temperature cyclers**. The cyclers offer well-to-well uniformity, improved cycling time and a gradient feature, which allows testing of up to 12 different PCR annealing temperatures in one run. Users can retrofit the Hot Top assembly to both non-gradient RoboCycler 40 models. Traditionally, a mineral oil or wax overlay has been used to prevent sample evaporation. This assembly works to prevent sample evaporation and condensation within the reaction tube, eliminating the need for messy and time-consuming oil or wax. The cycler unit and Hot Top assembly can be safely run overnight saving, resulting in additional time savings.

Reader Enquiry No. 100

Promega's TaqBead **'hot start' polymerase** is engineered to provide an easy, one-step method of performing hot start amplifications. Each wax bead contains 1.25 units of premeasured *Taq* DNA polymerase, rendering the enzyme inactive until



TaqBead one-step hot start amplification.

heated above 60 °C. The non-barrier format makes for easy pipetting after the reaction, and can be used without an oil overlay with heated-lid thermal cyclers. Hot starts are said to be especially useful when the amplification target is >1 kb. Performing a hot start has certain advantages over 'cold start' PCR, including reduction of nonspecific amplification products, enhanced low-copy detection and improved multiplex PCR.

Reader Enquiry No. 101

For the **detection of cytokine or adhesion molecule expression** R&D Systems Europe has launched the Ingenius range of matched PCR Primer Pairs for cytokine and adhesion molecule targets. The pairs

are supplied as an equimolar mixture of 5' and 3' primers with a positive control template provided with each set. Where specific information is available about the target, the primers have been designed to amplify across intron/exon junctions. Products amplified from cDNA or genomic DNA can therefore be distinguished by the size of the amplicon produced. Each matched pair has been carefully designed to minimize cross and self complementarity. Primers for related groups of molecules have also been tested to ensure that there is no cross-reactivity.

Reader Enquiry No. 102

Perkin-Elmer has introduced the ABI Prism 7700 **sequence detection system**, which is designed to allow the quantitative analysis of specific DNA sequences at increased speed. This fully automated system combines the company's chemistry with closed-tube, non-gel-based fluorescence detection to achieve high-throughput quantification of polymerase chain reaction as the PCR reaction proceeds. Using this system, researchers should be able to analyse up to 500 samples per day. This integrated system includes a built-in thermal cycler, a fluorogenic 5' nuclease assay, laser-induced fluorescence, charge-coupled device detection and PCR application software; no electrophoresis or post-PCR processing is necessary. According to the manufacturer, starting copy numbers ranging from hundreds to millions can be detected during the same PCR run. This should extend the dynamic range of PCR to approximately five orders of magnitude and eliminates the need to perform tedious and time consuming serial dilutions for quantitative studies. All reagents are added at one time and the fluorescent signal is detected through a closed tube.

Reader Enquiry No. 103

RT-PCR

Boehringer Mannheim has combined its single-tube mRNA capture with a 'set-up-and-walk-away' **RT-PCR system**. Using the mRNA Capture kit and Titan one-tube RT-PCR system, it is possible to isolate even low-abundance mRNAs from cells, tissue samples or total RNA preparations. The Titan system can perform single-tube RT-PCR, without changing buffers between reverse transcription and amplification. It has an optimized mixture



ABI Prism 7700 for sequence detection.

of the AMV reverse transcriptase and the high-fidelity Expand reagent to ensure high-specificity reverse transcription (at >50 °C), high-yield and high-fidelity amplification of fragments up to 6 kb. Like the mRNA Capture kit, the Titan system prevents sample loss and contamination.

Reader Enquiry No. 104

Amersham has launched a new RT-PCR kit for **first-strand cDNA synthesis**. This kit can be used to produce full length cDNA from low levels of mRNA and generates cDNA that can be used directly in RT-PCR experiments. Each kit is supplied with protocols and sufficient reagents to perform 100 reactions starting with as little as a stated 1 µg of purified mRNA or 5 pg of total RNA. Two types of primer are supplied with the kit, anchored oligo dT₂₅ primers and random hexamers, although specific oligo primers can also be used. This allows the most appropriate primer to be chosen, depending on the nature of the target mRNA. Anchored oligo dT₂₅ primers are designed to ensure that the resulting cDNA contains the minimum sequence from the poly(A) tail of the message, increasing the potential for full length synthesis. The kit can be used with either purified mRNA or with total RNA, if the target sequence is of medium to high abundance. Also included are control mRNA primers and PCR primers, which can be used to test the first strand cDNA reaction and the subsequent PCR amplification step.

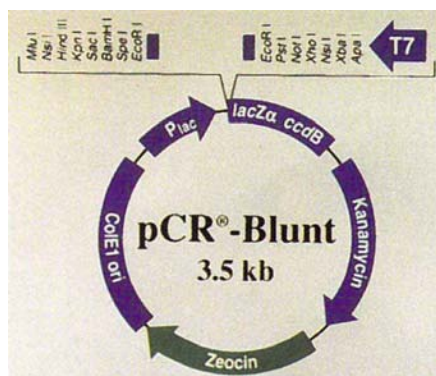
Reader Enquiry No. 105

Cloning

Invitrogen now offers the Zero Blunt PCR cloning kit for efficient, one-hour cloning of blunt-ended PCR products generated with proof-reading thermostable polymerases, such as Vent and *Pfu*. With this

PRODUCT REVIEW

system no post-PCR cleanup or enzymatic modifications are necessary, according to the manufacturer. In addition, there are no limitations to primer design, and no expensive restriction enzymes are required to reduce background. Zero Blunt uses the



Invitrogen's vector for blunt-end PCR.

pCR-Blunt vector, which allows for positive selection whereby only recombinants will grow. According to the manufacturer, 95 per cent of the colonies on the plate will be true recombinants. The kit includes linearized pCR-Blunt vector, PCR reagents, competent cells and controls.

Reader Enquiry No. 106

National Biosciences (NBI) has released the TC Blue cloning kit, which is a cloning system based on the T-A principle with an enhancement to improve the low efficiencies that may be associated with *Taq*-generated 3' G or 3' C overhangs. This design should assist researchers in cloning their PCR products with improved cloning efficiency, a greater assurance of positive transformants and more convenient post-cloning manipulations. The kit's multiple vector system is designed for different terminal extendase additions, including A, G and C, in all combinations. The 2.9-kb TC Blue vector should accommodate large inserts. For researchers lacking the time to subclone their PCR products, the company offers a service that will subclone the products into 3G Blue, TC Blue or any vector of interest; if preferred, NBI will also sequence the insert. The company provides a full complement of laboratory services for nucleic acid research, including custom DNA and custom RNA synthesis, phosphorothioate synthesis, oligonucleotide design services, custom DNA sequencing, gene synthesis, cloning, library screening and construction services.

Reader Enquiry No. 107

The new Fast-Link DNA ligation and screening kit developed by Epicentre is designed to reduce the time necessary to clone DNA fragments and identify recombinants. Ligation, transformation and

screening can be done in a stated 24 hours. The kit is designed to allow the ligation of any type of insert (cohesive ends, blunt ends or PCR fragments) and can be performed quickly with high transformation efficiency. According to the manufacturer, DNAs with blunt or cohesive ends can be ligated as little as five minutes; whereas ligation of PCR fragments with A-overhangs can be completed in one hour. Recombinants are said to be easily and quickly identified with the in-well lysis screening reagents and protocol, without the need to grow cultures or perform minipreps.

Reader Enquiry No. 108

Clean and neat PCR

DNA-OFF is a non-alkaline cleansing solution from CPG that is active against contaminating DNA. It is intended for use at PCR workstations, including counter tops, apparatus and pipettors. The presence of contaminant DNA at a PCR work area can result in DNA amplification artefacts. This solution contains a surfactant, as well as an agent to destroy DNA. The solution is stable, heat resistant, non-corrosive and does not contain carcinogenic compounds. Furthermore, it comes in an economical and convenient 500-ml size in ready-to-use form.

Reader Enquiry No. 109

Currently available from Anachem is a range of Bio 101 products, including the GeneClean kits used for the rapid removal and purification of DNA from solution or agarose gels. These kits are available in three sizes (200, 300 or 600 preps per kit) and have a 75–95 per cent yield per preparation. Equivalent kits are also available for RNA purification. Complementing the kits is the spin module, which can provide the user with a saving of about five minutes per prep versus non-spin kits. It should provide more consistent yield with low elution volumes and no matrix carryover.

Reader Enquiry No. 110

Intersep Filtration Systems has introduced the Fugisep centrifugal concentrator for the concentration and purification of large molecules. The range includes the Mini, Midi and Maxi models, offering flexible options for the concentration of biological molecules. The sealing technology is designed to ensure resistance to high gravitational forces (above 6,000g for Midi and Maxi, and 13,000g for Mini), thereby reducing sample processing time. Fitting into a standard microcentrifuge tube holder, the filter units are convenient for a range of applications, including protein concentration/diafiltration, DNA sample preparation (PCR) and fractionation of biological species. Inversion of filter tube and recentrifuga-

tion allows nearly complete recovery of concentrated product. With the Mini Fugisep, samples of up to 0.5 ml can be rapidly concentrated down to as low as 10 µl; Midi works with samples of up to 4.0 ml, which can be rapidly concentrated down to 40 µl; and Maxi handles samples of up to 15 ml that can be concentrated down to 50 µl. Another feature of the Fugisep range is that the concentration can be preselected (for example ×10, ×50, ×100) by introducing different starting volumes. The device comes in a variety of ultrafiltration molecular weight cut offs (4–100 K) and microlitre sizes (0.2 µm, 0.45 µm and 0.8 µm). A reverse osmosis membrane with a molecular weight cut off of 0.1 K is also included in the range.

Reader Enquiry No. 111

The Omni PCR workstation from Astec Microflow was designed to be a compact cabinet that provides protection for both the operator and the samples. The unit works on the principle that 70 per cent of the air drawn from the work area is recirculated. It passes through a main HEPA filter mounted at the top of the work area, at a velocity of 0.35 m s⁻¹. The self-balancing unit draws its make-up air through the front at a velocity stated to be 0.4 m s⁻¹ and discharges it through a HEPA filtered exhaust outlet. An ultraviolet decontamination system is interlocked with a microswitch on the door and cannot operate if the safety door is not fully closed. An adjustable timer allows samples, pipettes, and other items to be decontaminated. The cabinet's work surface, decontamination shelf and pipette holder are all removable and are constructed of 316 stainless steel. The sides and front door are polycarbonate.

Reader Enquiry No. 112

These notes are compiled by Brendan Horton from information provided by the manufacturers. For more details, fill in the reader service card bound inside the journal.

ADVERTISEMENT

READER ENQUIRY NO. 86